AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2015JD023733

Key Points:
- Evaluation of the water budget components in NCEP operational and research NLDAS-2 systems
- Focus on 12 National Weather Service's River Forecast Centers
- First using USGS HUC8 monthly runoff and gridded FLUXNET ET products to evaluate NLDAS-2 products

Correspondence to:
Y. Xia,
youlong.xia@noaa.gov

Citation:
Xia, Y., et al. (2016), Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems, *J. Geophys. Res. Atmos.*, *121*, 2750–2779, doi:10.1002/2015JD023733.

Received 29 MAY 2015
Accepted 8 MAR 2016
Accepted article online 12 MAR 2016
Published online 25 MAR 2016

Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems

Youlong Xia[1,2], Brian A. Cosgrove[3], Kenneth E. Mitchell[4], Christa D. Peters-Lidard[5], Michael B. Ek[1], Michael Brewer[6], David Mocko[5,7], Sujay V. Kumar[5,7], Helin Wei[1,2], Jesse Meng[1,2], and Lifeng Luo[8]

[1]Envronmental Modeling Center (EMC), National Centers for Environmental Prediction (NCEP), College Park, Maryland, USA, [2]I. M. Systems Group at NCEP/EMC, College Park, Maryland, USA, [3]National Water Center, National Weather Service, Silver Spring, Maryland, USA, [4]Prescient Weather Ltd, State College, Pennsylvania, USA, [5]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center (GSFC), Greenbelt, Maryland, USA, [6]National Centers for Environmental Information, NESDIS/NOAA, Ashville, North Carolina, USA, [7]SAIC at NASA/GSFC, Greenbelt, Maryland, USA, [8]Department of Geography, Michigan State University, East Lansing, Michigan, USA

Abstract The purpose of this study is to evaluate the components of the land surface water budget in the four land surface models (Noah, SAC-Sacramento Soil Moisture Accounting Model, (VIC) Variable Infiltration Capacity Model, and Mosaic) applied in the newly implemented National Centers for Environmental Prediction (NCEP) operational and research versions of the North American Land Data Assimilation System version 2 (NLDAS-2). This work focuses on monthly and annual components of the water budget over 12 National Weather Service (NWS) River Forecast Centers (RFCs). Monthly gridded FLUX Network (FLUXNET) evapotranspiration (ET) from the Max-Planck Institute (MPI) of Germany, U.S. Geological Survey (USGS) total runoff (Q), changes in total water storage (dS/dt, derived as a residual by utilizing MPI ET and USGS Q in the water balance equation), and Gravity Recovery and Climate Experiment (GRACE) observed total water storage anomaly (TWSA) and change (TWSC) are used as reference data sets. Compared to these ET and Q benchmarks, Mosaic and SAC (Noah and VIC) in the operational NLDAS-2 overestimate (underestimate) mean annual reference ET and underestimate (overestimate) mean annual reference Q. The multimodel ensemble mean (MME) is closer to the mean annual reference ET and Q. An anomaly correlation (AC) analysis shows good AC values for simulated monthly mean Q and dS/dt but significantly smaller AC values for simulated ET. Upgraded versions of the models utilized in the research side of NLDAS-2 yield largely improved performance in the simulation of these mean annual and monthly water component diagnostics. These results demonstrate that the three intertwined efforts of improving (1) the scientific understanding of parameterization of land surface processes, (2) the spatial and temporal extent of systematic validation of land surface processes, and (3) the engineering-oriented aspects such as parameter calibration and optimization are key to substantially improving product quality in various land data assimilation systems.

1. Introduction

Since the multi-institution North American Land Data Assimilation System (NLDAS) was initiated in 2000, it has been conducted in two phases. Phase 1 initiatives spanned from 2000 to 2005 (NLDAS-1 [*Mitchell et al.*, 2004]). The objectives of Phase 1 were to establish the NLDAS configuration (including collection of soil and vegetation data), selection of land surface models (LSMs), generation of surface forcing data sets, and retrospective model runs for a 3 year period (from October 1996 to September 1999), with evaluation/validation of model output. Phase 2 is an extension of Phase 1. The purpose of NLDAS is to support numerous applications for researchers and operational users in both the land modeling community and the water resources management community, including drought monitoring. The Phase 2 initiatives included (A) much longer retrospective simulations starting from January 1979 to 2009 executed during the 2006 to 2009 time frame, followed by (B) the first North American Land Data Assimilation System version 2 (NLDAS-2) real-time quasi-operational simulations, including extensive product evaluation and pilot product application from 2010 to 2012 (NLDAS-2 [*Ek et al.*, 2011; *Xia et al.*, 2012a, 2012b, 2013]). In August 2014, NLDAS-2 was formally implemented into official National Centers for Environmental Prediction (NCEP) operations (http://www.nco.ncep.noaa.gov/pmb/products/nldas).

NCEP/EMC (Environmental Modeling Center) now maintains two NLDAS-2 versions: (1) a version that mimics the NCEP operational version and (2) a research version. In the operational NLDAS-2, the four land surface models (LSMs) of Mosaic, Noah, SAC, and VIC are the same versions as used in NLDAS-1, except for the Noah LSM. In the latter, the snowpack physics were enhanced to improve water and energy flux simulation in the cold season, following the work of *Livneh et al*. [2010], and some parameters were modified to improve water and energy flux simulation in the warm season [*Wei et al*., 2013]. In the research version of NLDAS-2, all four LSMs except for Mosaic were upgraded through either tuning model parameters (i.e., Noah-I [*Xia et al*., 2014c]; VIC4.05 [*Troy et al*., 2008]) or using an improved formulation of potential evapotranspiration (PET) (i.e., SAC-Clim [*Xia et al*., 2012a, 2012b]). These upgraded LSMs are not implemented yet in the NCEP operational NLDAS-2 as they need more assessment to evaluate their performance, such as the new assessments provided in this present study.

Additionally in the NLDAS-2 research suite, the NCEP/EMC NLDAS team is collaborating with its partner—the NASA/GSFC (Goddard Space Flight Center) Hydrological Sciences Laboratory to develop the third-phase NLDAS system (the future NLDAS-3), which will feature (1) a LIS-based (Land Information System) [*Kumar et al*., 2006; *Peters-Lidard et al*., 2007] ensemble Kalman filter (EnKf) assimilation capability and (2) replacement of the Mosaic LSM [*Koster and Suarez*, 1994] with the NASA Catchment LSM [*Koster et al*., 2000; *Ducharne et al*., 2000], along with upgrades of the other three LSMs (Noah, SAC, and VIC) to their latest versions. The preliminary results from this pilot NLDAS-3 have shown that both the upgrade of the LSMs and the addition of a full-fledged land data assimilation capability can improve total runoff and soil moisture simulation [*Kumar et al*., 2014]. The collaboration is an ongoing effort and is still underway.

The hydrometeorological products produced by *two LSMs* in the *operational* NLDAS-2 (Mosaic and Noah) and by *two LSMs* in *the research* NLDAS-2 (SAC-Clim and VIC4.0.5) have been evaluated against many in situ observations and satellite retrievals, such as U.S. Geological Survey (USGS) streamflow [*Xia et al*., 2012b], soil temperature [*Xia et al*., 2013], soil moisture [*Xia et al*., 2014a], and evapotranspiration [*Xia et al*., 2014b]. However, the hydrometeorological products generated by the *other two LSMs* in the *operational* NLDAS-2 (i.e., SAC and VIC) and the one other upgraded LSM in the *research* NLDAS-2 (i.e., Noah-I) have not yet been comprehensively evaluated. One exception is that the soil moisture simulations from all four LSMs in the operational NLDAS-2 have been evaluated against in situ soil moisture observations by *Xia et al*. [2015a]. We here remind the reader that the version of Mosaic in NLDAS-2 is identical to that in NLDAS-1.

The purpose of the present paper is to comprehensively evaluate the products encompassing all components of the surface water budget from all four LSMs in both the operational and research NLDAS-2 at the basin scale of the 12 NWS River Forecast Centers. Following the operational NLDAS-2 assessment on the larger continental scale in *Xia et al*. [2012a, 2012b], we have chosen to first evaluate the surface water budget components (precipitation, total runoff, evapotranspiration, and total water storage change). Second, a recent companion paper has evaluated the surface *energy* balance components of all four LSMs in the operational and research NLDAS-2 [*Xia et al*., 2015b].

NLDAS-2 gridded precipitation fields are derived mainly from gauge observations. As a basis of comparison, these precipitation fields will be compared with the newly released monthly gridded precipitation data set [*Vose et al*., 2012; 2014] of the National Climate Data Center (NCDC) (Note: NCDC was renamed recently as the "National Centers for Environmental Information" (NCEI) but is still cited as NCDC throughout this paper to allow, for example, retention of the NCDC label in figures.). If NLDAS-2 precipitation (P) is considered to be an observed variable, then once observation-based fields of both monthly total runoff (Q) and monthly evapotranspiration (ET) are obtained, the total monthly water storage change (dS/dt) can be derived as the residual (P-Q-ET) of the surface water balance. Therefore, in this study, producing observation-based Q and ET benchmark fields are two central thrusts, the approaches for which are described in section 3.1.

This paper is organized as follows. A brief description of each LSM is presented next in section 2. The data sets used in this study are described and explained in section 3. The mean annual climatology and the annual cycle of monthly values from the four LSMs and their multimodel ensemble mean (MME), as well as monthly variations of total water storage anomaly/change, are evaluated in the operational versions and research versions in sections 4 and 5, respectively. Finally, the results of the study are summarized and the conclusions are presented in section 6, and the future pathway is given in section 7.

Table 1a. Primary Attributes of the Four NLDAS-2 Land Surface Models, Operational NLDAS-2 Version (Modified From *Mitchell et al.* [2004])

Model	Mosaic	Noah (Version 2.8)	VIC (Version 4.0.3)	SAC
Input surface forcing	7 fields	7 fields	7 fields	*P*, Noah PET, 2-m air T
Energy balance	Yes	Yes	Yes	n/a
Water balance	Yes	Yes	Yes	Yes
Number of model soil layers	3	4	3	2 storages
Soil layer depths (m)	0.1, 0.3 1.6	0.1, 0.3, 0.6, 1.0	Top layer 0.1 m, other layers vary with grid cell	Vary with grid cell
Tiling of vegetation	Yes	No	Yes	No
Number of snow model layers	1	1	2	1
Frozen soil: thermal	No	Yes	Disabled	n/a
Frozen soil: hydraulics	partial	Yes	Disabled	No
Soil temperature profile	No	Yes	Yes	n/a
Soil thermodynamics	Partial force-restore	Heat conduction equation	Heat conduction equation modified	n/a
Soil water drainage	Yes	Yes	Yes	Yes
Soil water vertical diffusion	Yes	Yes	No	No
Explicit vegetation	Yes	Yes	Yes	No
Canopy resistance	*Sellers et al.* [1986]	*Jarvis* [1976]	*Jarvis* [1976]	n/a
Root depth	0.4 m	1 or 2 m	1.35–3 m	n/a
Note	Same as Mosaic used in NLDAS-1	Winter and summer physics update [*Livneh et al.*, 2010; *Wei et al.*, 2013]	Same as VIC used in NLDAS-1	Same as SAC used in NLDAS-1

2. Operational and Research NLDAS-2 Systems

As noted earlier, the EMC NLDAS team maintains two versions of the NLDAS-2 system: the NCEP operational version and a research version. The operational version is run and maintained by NCEP Central Operations as part of NCEP Production Suite. Its purpose is to provide reliable 24/7/365 near-real-time hydrometeorological products to support operational U.S. Drought Monitor analysis and prediction tasks, as well as the operational water- resource analysis and prediction tasks of other federal, state, and municipal governmental agencies, universities, and the private sector. The *research* version is run and maintained by the EMC NLDAS team. Its purpose is research and development (R&D) to achieve next-generation improvements to NLDAS-2 LSMs and their hydrometeorological products. Viable R&D requires execution and reexecution of long respective periods, which in EMC have recently spanned 1979–2012. The advantage of the research version is that upgraded model physics (to address known existing model biases and simulation shortcomings) can be relatively quickly tested and assessed. Their performance can be evaluated against either in situ observations and/or satellite retrievals [*Xia et al.*, 2012a, 2012b; *Xia et al.*, 2014a] and/or older-generation NLDAS simulations, by which means known model shortcomings can often be reduced significantly.

Tables 1a and 1b compare the attributes of the four LSMs (operational versus research) in NLDAS-2. These LSMs feature a good cross section of developmental legacies. The Mosaic [*Koster and Suarez*, 1994] and Noah [*Ek et al.*, 2003] LSMs emerged from the surface-vegetation-atmosphere transfer (SVAT) setting of coupled atmospheric land models in weather and climate prediction models and therefore treat both the surface energy balance and surface water balance. LSMs with this type of legacy, which emerged with relatively little focus on calibration, are typically "grid point" models (also known as "distributed" models) that execute on a well-defined computational grid. The SAC [*Burnash et al.*, 1973] and VIC [*Liang et al.*, 1994] LSMs grew out of the hydrology community as uncoupled hydrologic models with considerable calibration. In recent years, however, VIC has evolved toward an SVAT-like LSM [*Wood et al.*, 1997], albeit still with substantial calibration (typically targeting observed *Q*), which was used in NLDAS-1 [*Mitchell et al.*, 2004].

Table 1b. Primary Attributes of the Four NLDAS-2 Land Surface Models, Research NLDAS-2 Version

Models	Primary Attributes
Mosaic	Same as Mosaic used in operational version
Noah-I (Noah-Interim)	Modify CH constraint only for snow-covered area [*Xia et al.*, 2014c]
VIC4.0.5 (version 4.0.5)	Tuned soil parameters [*Troy et al.*, 2008]
SAC-Clim	Seasonally varied monthly PET climatology [*Xia et al.*, 2012a]

Therefore, Mosaic, Noah, and VIC can simulate both the surface water balance (including snowpack and soil moisture in several soil layers) and the surface energy balance (including land skin temperature and soil temperature in several soil layers). SAC was originally developed as a lumped conceptual hydrological model, is typically highly calibrated for specific small catchments, and is used operationally in NWS RFCs. For purposes of executing in NLDAS-1 alongside the other three LSMs above, the NWS Office of Hydrologic Development (now the National Water Center—NWC) developed a semidistributed version of the SAC model (applied on the NLDAS grid).

As shown in Tables 1a and 1b, the versions of Mosaic, VIC, and SAC used in the operational NLDAS-2 are the same versions as are used in the R&D-oriented NLDAS-1. In this regard, the Noah LSM is an exception, with Noah version 2.7.1 used in NLDAS-1, while the upgraded Noah version 2.8 is used in the operational NLDAS-2. Compared to Noah 2.7.1, Noah 2.8 includes (for the cold season) enhanced snowpack physical processes, such as adding snow aging [*Livneh et al*., 2010] and some additional tuning of some model parameters (e.g., maximum snow albedo and aerodynamic conductance), and enhance model physical processes for warm season (e.g., adding seasonally varied leaf area index, adding effect of seasonal root uptake activity to transpiration calculation [*Wei et al*., 2013]). For the *research* NLDAS-2, Noah 2.8 was further upgraded to an "intermediate" Noah version (Noah-I) by adjusting the aerodynamic conductance for snow-free regions to help reduce (1) a negative ET bias in north central and northeastern regions of the continental United States (CONUS) and (2) a negative land skin temperature and soil temperature bias in cold regions and seasons [*Xia et al*., 2014c]. In their research versions, SAC was changed to use the observed monthly potential evapotranspiration (PET) climatology to replace the bias-corrected Noah PET (SAC-Clim), and VIC was upgraded to VIC4.0.5 by tuning its soil parameters [*Troy et al*., 2008]. More details about the LSMs and setup for both NLDAS-1 and NLDAS-2 are given in *Mitchell et al*. [2004] and *Xia et al*. [2012a].

It should be noted that all four land surface models and their upgraded versions used in this study do not resolve the influences of irrigation (IR) and groundwater (GW) on ET. The RFCs for which the IR and GW processes are anticipated to have significant influence on ET are ABRFC, CNRFC, MBRFC, NCRFC, and WGRFC. While this lack of the IR and GW processes may affect our results analyzed here, insight can be gained by examining preexisting comparisons conducted with the operational version of Noah (Noah2.8) and the research version of VIC (VIC4.0.5) in the NLDAS-2 system. These models have been evaluated against in situ observations, Max Planck Institute (MPI) gridded ET, and Gravity Recovery and Climate Experiment (GRACE)-observed total water storage anomalies and have also been compared with Noah-MP (Noah-Multi-Physics) [*Niu et al*., 2011; *Yang et al*., 2011] and CLM4 (Community Land Model version 4) [*Gent et al*., 2010; *Lawrence et al*., 2012] output. The latter is significant as both Noah-MP and CLM4 include groundwater modules and are able to simulate variations in the ground water table and the exchange of water between the deep soil and groundwater stores. These intercomparisons did not yield substantial ET differences between Noah-MP/CLM4 and Noah/VIC4.0.5, although some regionally and seasonally varying differences do exist [*Cai et al*., 2014]. Therefore, it is speculated that the lack of IR and GW modules in the models of this study may exert only a modest impact across arid and semiarid RFCs such as CBRFC, CNRFC, ABRFC, and WGRFC in this analysis. Future investigation will be needed to further define these impacts.

3. Data Sets and Method

3.1. Data Sets Used

The data sets used in this study include simulated products from both the operational and research NLDAS-2 (NCEP website: http://www.emc.ncep.noaa.gov/mmb/nldas; NASA website: http://ldas.gsfc.nasa.gov/nldas/) and observations and observation-based reference products from different data sources. Over the CONUS, NLDAS-2 precipitation is derived from the Climate Prediction Center (CPC) unified gauge-based precipitation analysis with monthly Parameter-elevation Regressions on Independent Slopes Model (PRISM) [*Daly et al*., 1994] adjustments for orographic impacts on precipitation. In areas where these data are unavailable, North American Regional Reanalysis (NARR) precipitation is used instead. Since the NARR assimilates precipitation gauge data, the merged CPC-NARR-based precipitation forcing field is relatively seamless [*Mesinger et al*., 2006].

The 2 m air temperature from the 32 km NARR is bilinearly interpolated to the 1/8° NLDAS2 grid and adjusted using an elevation and constant lapse rate (6.5°C/km) approach. The air temperature is used to

separate total precipitation into snowfall and rainfall using a 0.0°C threshold value. This partitioning affects the simulation of snow water equivalent, which is used to evaluate the GRACE-observed total water storage anomaly (TWSA).

The USGS-observed streamflow at 986 basins was used in our previous streamflow validation [*Xia et al*., 2012b]. These 986 small- to medium-sized basins, chosen for being deemed relatively free of human control, cover only one third of the continental United States (CONUS). Most of them are located in the eastern half or near the western coast of the CONUS, with relatively few basins in drier interior regions west of the Mississippi. Therefore, streamflow validation was performed in only these limited regions.

Recently, *Velpuri et al*. [2013] used USGS HUC-based (Hydrological Unit Code) monthly runoff to evaluate mean annual ET via the water balance equation and thereby obtained quite reasonable results (using the common assumption that on an annual basis, the water storage change term of the water balance equation is negligible). In Velpuri et al., the entire CONUS is divided into 21 major regions (HUC2) composed of 222 subregions (HUC4), which are further divided into smaller basins (HUC6) and subbasins (HUC8). The boundaries of these units are defined in terms of topographic river basin divides and subbasins. The eight-digit hydrologic unit codes are composed of two digits each for region, subregion, basin, and subbasin. The approximate mean area of regions, subregions, basins, and subbasins, respectively, are 500,000; 50,000; 25,000; and 4000 km^2 [*Velpuri et al*., 2013].

The estimates of monthly runoff for each HUC8 were derived by the USGS by combining (1) historical flow data from stream gauges, (2) drainage area of the basins upstream of the stream gauges, and (3) the boundaries of the HUC8. The latter USGS runoff product for the HUC8s, used herein, provide an important new opportunity to evaluate the simulated total runoff of each LSM in NLDAS-2, though keeping in mind that the nominal HUC8 area is much larger than an NLDAS grid box (~200 km^2). The HUC8 data were derived from the comprehensive USGS National Water Information Service (NWIS) gauge observation data, the respective drainage basin boundaries of the streamflow gauges, and the boundaries of each eight-digit hydrologic unit.

The monthly runoff is the accumulated time series of flow per unit area calculated for each HUC8 subbasin. For each HUC8 subbasin, multiple NWIS gauge stations located within or downstream of the HUC8 were used to estimate the runoff generated locally at each HUC8. The contributing drainage areas (both gauge-to-HUC8 and HUC8-to-gauge) were converted as weighting factors to merge runoff time series from all stations. As a result, gauges with drainage coverages most similar to that of the particular HUC8 received the highest weights. Therefore, the influence of highly regulated gauge stations (usually with large drainage coverages across multiple HUC8s) was reduced. This approach may effectively merge streamflow observations from multiple gauge stations as a consistent areal HUC8 runoff measurement, which can be considered as a close surrogate to the natural runoff [*Ashfaq et al*., 2013; *Oubeidillah et al*., 2014].

For the ET evaluation, we use the Max Planck Institute (MPI) flux data set of *Jung et al*. [2009], which was generated by using a multitree ensemble (MTE) method to synthesize FLUXNET tower data [*Baldocchi et al*., 2001] with meteorological forcings and vegetation information from interpolated station and satellite data. The MTE method produces a global, monthly, 1/2° resolution estimate of land ET from 1982 to 2008 [https://www.bgc-jena.mpg.de/geodb/projects/Data.php]. The gridded MTE FLUXNET ET data have been used as the key reference ET data set for evaluating multiple global ET products produced using four different categories of techniques: (1) observations-based diagnostic data sets; (2) observationally driven offline land surface model products; (3) atmospheric reanalyses (which include a coupled LSM); and (4) Intergovernmental Panel on Climate Change AR4 simulations from 11 general circulation models [*Jiménez et al*., 2011; *Mueller et al*., 2011; *Velpuri et al*., 2013].

Some NLDAS-2 ET products have been evaluated against this reference data set [*Peters-Lidard et al*., 2011; *Cai et al*., 2014]. To reduce the scale mismatch issue in these evaluations, *Peters-Lidard et al*. [2011] used an upscaling method to aggregate NLDAS-2 ET products from 0.125° to 0.5°, and *Cai et al*. [2014] used a downscaling method (i.e., water budget algorithm—a bilinear interpolation algorithm that conserves water within the interpolation procedure [*Sharif and Ogden*, 2014]) to downscale the MTE FLUXNET ET product from 0.5° to 0.125°. We recognize that the former upscaling method is more preferable (for the purpose of evaluating products) than the latter downscaling method, wherein the spatial interpolation error is not trivial.

It should be noted that the MPI ET product is not an observed product (e.g., FLUXNET eddy covariance measurements) but rather a model-based product. FLUXNET observations of carbon dioxide, water, and energy fluxes are first upscaled to the global scale using the MTE method. The MTE is trained to predict

site-level gross primary productivity, terrestrial ecosystem respiration, net ecosystem exchange, latent heat flux, and sensible heat flux based on remote sensing indices, climate and meteorological data, and information on land use. The trained MTEs are then applied to generate global flux fields at a $0.5° \times 0.5°$ spatial resolution and a monthly temporal resolution from 1982 to 2008 [*Jung et al*., 2009; *Jung et al*., 2011]. Cross-validation analyses revealed good performance of the MTE in predicting inter-site flux variability with modeling efficiencies between 0.64 and 0.84. Yet this product is still hampered by a variety of challenges, such as a limited ability to account for the effects of disturbance and/or site history and lagged environmental effects. Moreover, this product has very low interannual variability when compared to the FLUXNET tower observations [*Baldocchi et al*., 2001].

The latest release of the GRACE-observed total water storage anomaly—TWSA (RL05 [*Landerer and Swenson*, 2012]) is obtained from the GRACE Tellus website (ftp://podaac-ftp.jpl.nasa.gov/allData/tellus/L3/land_mass/RL05/netcdf/). The RL05 includes three products processed at three different centers: the Center for Space Research (CSR) and the Jet Propulsion Laboratory (JPL) in the U.S. and GeoForschungZentrum (GFZ) in Germany. The anomalies of each of the three products are calculated relative to each product's own climatology (i.e., with a 2004–2009 time mean). The products feature a spatial resolution of 1° and a monthly time step. The RL05 product, spanning December 2002 to January 2015, contains a total of 146 months of data with 10 months of missing data (e.g., June 2003; May and October 2012; March, August, and September 2013; and February and December 2014). To reduce attenuated surface mass variations at small spatial scales due to the sampling and postprocessing of GRACE-based observations, and to provide uncertainty estimates for users, the website also provides gridded scaling factor and error estimates, including measurement and leakage errors. As suggested by the GRACE Tellus website (http://grace.jpl.nasa.gov/data/choosing-a-solution/) and recent work from *Sakumura et al*. [2014], a simple mathematical average of the CSR, GFZ, and JPL products is the most effective approach to reduce the noise from the three different products. The TWSA is calculated as the product of the simple average and a scale factor. The total water storage change (TWSC) is derived from the TWSA using a centered difference approximation $[(\mathrm{TWSA}\ (t+1) - \mathrm{TWSA}\ (t-1))/2]$, as it can more efficiently reduce noise than simple forward difference approximation [*Long et al*., 2014]. In the TWSC processing procedure, missing TWSA data are filled using a temporal linear interpolation algorithm. The uncertainty of the TWSA is estimated by the total error calculated from the measurement error and leakage error [*Swenson et al*., 2006; *Wahr et al*., 2006]. The uncertainty in GRACE-derived TWSC is computed from the uncertainty in GRACE-observed TWSA for surrounding months added in quadrature. The GRACE-observed TWSA and TWSC have been widely used to evaluate land surface model products [*Long et al*., 2013; *Getirana et al*., 2014] and groundwater monitoring [*Strassberg et al*., 2009].

Due to the different spatial scales characterizing the data sets analyzed in this work (which include 0.0416° for NCDC precipitation, USGS Q (~4000 km^2), MTE FLUXNET ET (0.5°), GRACE-observed TWSA and TWSC (1°), NLDAS-2 precipitation, Q, ET, total column soil moisture, snow water equivalent (SWE), and canopy water storage products (0.125°)), a direct comparison between NLDAS-2 simulated products and observation-based Q, ET, TWSA, and TWSC may suffer from scale mismatch problems. As we prefer upscaling over downscaling to reduce scale mismatch issues, we apply spatial averaging across the overall basin of responsibility of each NWS RFC. RFC basins are an appealing spatial averaging choice, because the hydroclimatology across a given RFC basin is reasonably self-similar spatially. An analogous method has been used for ET evaluation in NLDAS-1 and NLDAS-2 [*Robock et al*., 2003; *Mo et al*., 2011; *Peters-Lidard et al*., 2011; *Xia et al*., 2014b].

From the NCEP/EMC NLDAS-2 ftp site (ftp://ldas.ncep.noaa.gov/nldas2), we obtained the 27 year (1982–2008) *hourly* NLDAS-2 precipitation, total runoff, and evapotranspiration for both the operational and research versions of the four LSMs in NLDAS-2 (for Mosaic, the operational and research data sets are the same). Monthly PE climatology fields used by SAC-Clim as forcing input were obtained from the NWS/NWC. It should be noted that with the exception of PE for SAC-Clim, the same forcing data including precipitation are used for both the operational and research NLDAS-2 systems.

The 27 year *monthly* values were computed from the hourly values for the continental United States. Per *Vose et al*. [2014], we obtained the corresponding 27 year, monthly NCDC precipitation fields at 0.0416° resolution, and we upscaled the latter precipitation fields to NLDAS-2 0.125° resolution using the water budget method. We followed *Cai et al*. [2014], who used a water budget method to regrid the FLUXNET ET fields from 0.5° to 0.125° over the CONUS, and we then aggregated the interpolated ET for each RFC. We then compared

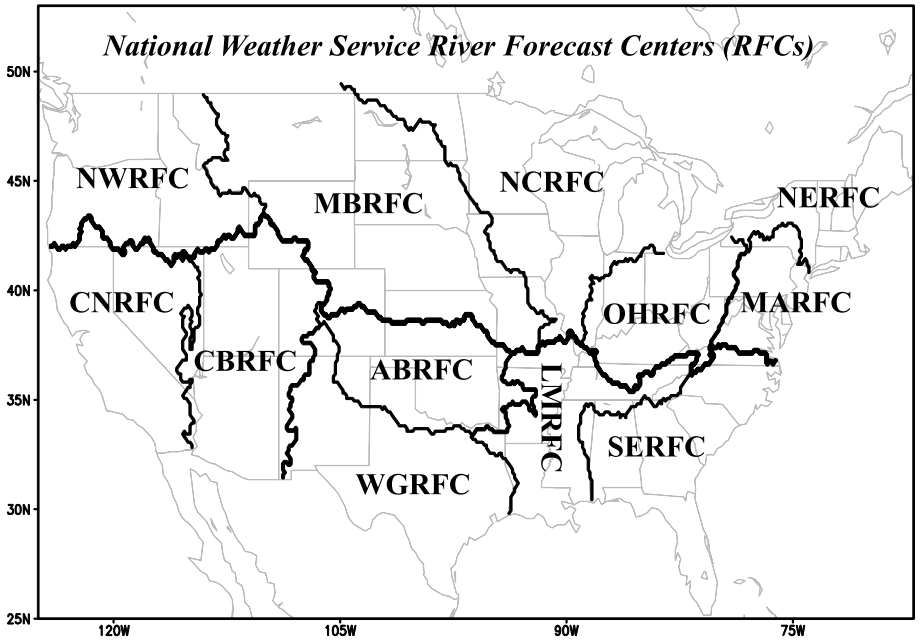

Figure 1. Location and area of each of the 12 National Weather Service (NWS) River Forecast Centers (RFCs) within the continental United States (CONUS).

RFC-averaged reference ET and NLDAS-2 simulated ET. We obtained the 27 year USGS monthly runoff estimates (Q) for HUC8 from the USGS Waterwatch website (http://waterwatch.usgs.gov/index.php?id=romap3). The HUC8-index text file was created using NLDAS-2 grid cells. The monthly HUC8 Q values were overlaid on the NLDAS-2 grid cells over CONUS. The 27 year monthly total water storage change (TWSC) was derived from the difference between NLDAS-2 monthly precipitation (P) and the sum of USGS total runoff (Q) and MPI FLUXNET ET. Lastly, we used a mask (defined on the NLDAS-2 computational grid, Figure 1) delineating each RFC basin to calculate the spatially averaged basin-mean monthly P, Q, ET, and dS/dt for each of the 12 RFCs.

To compare GRACE-observed TWSA and TWSC for a 12 year (2003–2014) period, monthly total column soil moisture, snow water equivalent, canopy water storage, P, Q, and ET from the four models in the operational NLDAS-2 system were downloaded from the NCEP/EMC NLDAS-2 website. Products from the research NLDAS-2 system were not evaluated as they do not cover this full period. It should be noted that the LSM-simulated total water storage only includes total column soil moisture, snow water equivalent, and canopy water storage and lacks ground water and explicit reservoir storage. Therefore, some differences can be expected in the following comparisons. In addition, as there is no explicit representation of rivers, lakes, or reservoirs in NLDAS-2, water storage in these water bodies is not included. This exclusion is not realistic; however, it is reasonable because these water bodies are excluded from the calculation of the water balance in the LSMs as well.

These Q, TWSA/TWSC, and ET fields have been used as key reference data sets in various previous studies [*Jiménez et al.*, 2011; *Mueller et al.*, 2011; *Peters-Lidard et al.*, 2011; *Long et al.*, 2013; *Velpuri and Senay*, 2013; *Velpuri et al.*, 2013; *Cai et al.*, 2014; *Getirana et al.*, 2014; *Long et al.*, 2014; *Oubeidillah et al.*, 2014].

3.2. Evaluation Method

The main thrusts of this evaluation are to compare the observed and simulated mean annual climatology, as well as the seasonal cycle of monthly values for ET, Q, and dS/dt for both the operational and research versions of NLDAS-2, and to compare the monthly variation of TWSA/TWSC in the operational version of NLDAS-2. Statistics such as correlation (R), anomaly correlation (AC), bias (Bias), standard deviation (Sigma), and root-mean-square error/deviation (RMSD) were used for comparison. In order to test if two dependent correlations are significantly different, we used the three correlation coefficients and their associated sample sizes to compute the probability value and the z score (Steiger's z test [*Steiger*, 1980; *Saville*, 1990]). A probability

Table 2. RFC Names and 27 Year (1982–2008) Climatology of Water Budget Components (Unit: mm/Year) Calculated From Several Sources[a]

Label	RFC Name	NCDC *P* (*P*1) (mm)	NLDAS-2 *P* (*P*2) (mm)	USGS *Q* (mm)	FLUXNET ET (mm)	$\frac{P1}{P2}$ (-)	$\frac{P2}{Q+ET}$ (-)
CBRFC	Colorado	358.2	353.6	53.1	270.5	1.01	1.09
CNRFC	California-Nevada	472.6	459.4	164.3	324.7	1.03	0.94
WGRFC	West Gulf	634.9	627.2	68.3	515.1	1.01	1.08
MBRFC	Missouri	541.7	550.2	73.2	437.4	0.99	1.08
ABRFC	Arkansas	751.6	754.3	105.6	546.8	1.00	1.16
NCRFC	North Central	804.6	808.5	242.8	517.1	1.00	1.06
NWRFC	Northwest	810.0	804.2	492.1	399.7	1.01	0.90
MARFC	Mid-Atlantic	1103.3	1093.5	468.1	597.8	1.01	1.03
SERFC	Southeast	1291.2	1278.6	392.2	786.6	1.01	1.09
NERFC	Northeast	1137.0	1131.6	637.8	482.0	1.01	1.01
LMRFC	Lower Mississippi	1384.6	1348.5	486.6	777.0	1.03	1.07
OHRFC	Ohio	1128.4	1103.5	466.5	641.3	1.02	1.00

[a]The RFCs are listed in order of increasing value of *P*/PET as shown in Table 3, varying from dry climate to wet climate.

value of less than 0.05 indicates that the two correlation coefficients are significantly different from each other at the 95% confidence level for a two-tailed test. If the anomaly correlation in the research version is larger than that from the operational version, this indicates an improvement in LSM simulation performance. Otherwise, it indicates a deterioration in performance.

4. Evaluation of Water Budget Components for Operational NLDAS-2

4.1. Mean Annual Climatology Analysis

In evaluating the 27 year mean annual water budget, the magnitude of the 27 year change in water storage (d*S*/d*t*) is negligible compared to the remaining terms in the water balance equation, which thus can be approximated as

$$P = \mathrm{ET} + Q \tag{1}$$

where *P* is total annual precipitation, ET is total annual evapotranspiration, and *Q* is total annual runoff. When the observed *P* and *Q* are available, water-balance-based ET can be derived from equation (1) by using *P*-*Q*. This approach assumes no long-term change in groundwater or surface water storage as the four models did not include groundwater or irrigation submodels. While this assumption may be an oversimplification in areas of the midwestern CONUS where irrigation is prevalent, it has been used in other long-term mean annual hydroclimatology analyses including *Hobbins et al.* [2001] and *Sankarasubramanian and Vogel* [2002]. Future model enhancements will allow for a more sophisticated treatment of the irrigation issue. The corresponding RFC names, mean annual NCDC precipitation, NLDAS-2 precipitation, FLUXNET ET, USGS *Q*, ratio between NCDC and NLDAS-2 precipitation, and ratio between NLDAS-2 precipitation and sum of FLUXNET ET and USGS *Q* are given in Table 2.

In Table 2, the mean annual precipitation varies from 353.6 mm/year at CBRFC to 1348.5 mm/year at LMRFC. Mean annual ET varies from 270.5 mm/year at CBRFC to 786.6 mm/year at SERFC, and mean annual *Q* varies from 53.1 mm/year at CBRFC to 637.8 mm/year at NERFC. The results show that NCDC and NLDAS-2 precipitation amounts are very comparable at the 12 RFCs (their difference is smaller than 3%) showing that NLDAS-2 precipitation is quite robust as an observation-based precipitation analysis on RFC basin spatial scales. The ratio between NLDAS-2 precipitation and (ET + *Q*) varies from 0.90 at NWRFC to 1.16 at ABRFC. Thus, the difference between NLDAS-2 precipitation and (ET + *Q*) is less than 10% for all 12 RFCs except for ABRFC, indicating a suitably small error in the water balance of equation (1).

Figure 2 presents the spatial distribution of 27 year mean annual NCDC precipitation (Figure 2a), NLDAS-2 precipitation (Figure 2b), FLUXNET ET (Figure 2c), USGS *Q* (Figure 2d), Mosaic ET (Figure 2e), Noah ET (Figure 2f), SAC ET (Figure 2g), VIC ET (Figure 2h), Mosaic *Q* (Figure 2i), Noah *Q* (Figure 2j), SAC *Q* (Figure 2k), and VIC *Q* (Figure 2l). NCDC and NLDAS-2 mean annual precipitation have very similar spatial distributions and small differences, consistent with their ratio shown in Table 2. However, the differences between NLDAS-2 and NCDC may not be trivial for some regions and individual months as different precipitation data sources and interpolation algorithms are used. NLDAS-2 precipitation is generated from a daily CPC gauge-based precipitation analysis produced by an optimal interpolation algorithm [*Gandin*, 1963], and NCDC monthly precipitation is generated by a thin-plate spine method [*Hutchinson*, 1995].

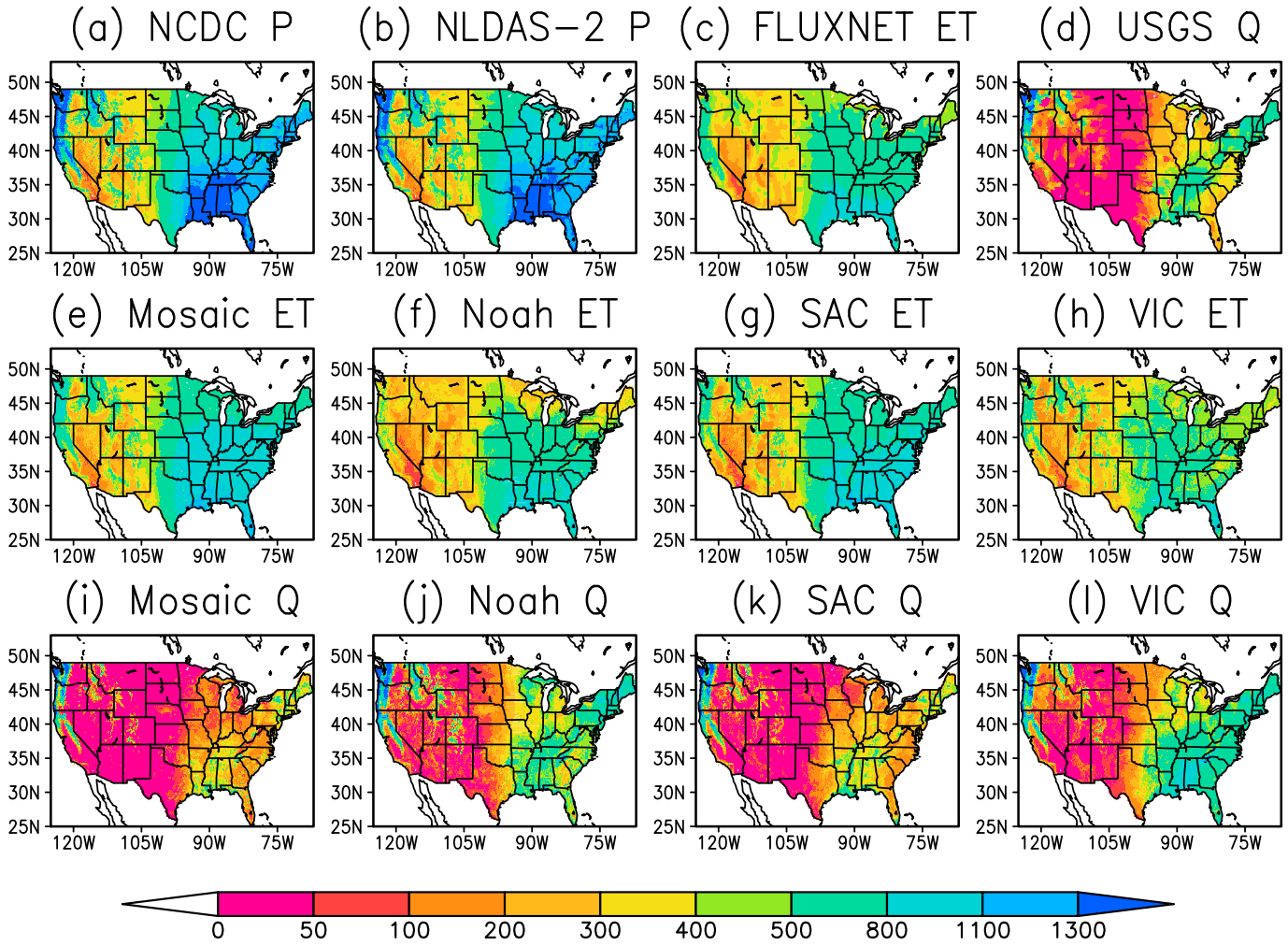

Figure 2. (top row) Mean annual precipitation (*P*) from (a) NCDC and (b) NLDAS-2, (c) mean annual evapotranspiration (ET) from MTE FLUXNET, and (d) mean annual total runoff (*Q*) from USGS. (middle row) mean annual ET in the operational NLDAS-2 from (e) Mosaic, (f) Noah, (g) SAC, and (h) VIC. (bottom row) Mean annual *Q* in the operational NLDAS-2 from (i) Mosaic, (j) Noah, (k) SAC, and (l) VIC. All results are for January 1982 to December 2008 (units: mm/year).

All four LSMs can broadly capture the spatial distribution of ET and *Q* observations, although Mosaic and SAC in general overestimate FLUXNET ET and underestimate the USGS *Q*. In contrast to Mosaic and SAC, Noah and VIC underestimate the FLUXNET ET and overestimate the USGS *Q*. In particular, at SERFC VIC largely underestimates the FLUXNET ET and largely overestimates the USGS *Q*, consistent with the streamflow results in NLDAS-1 [*Lohmann et al*., 2004; *Mitchell et al*., 2004].

Mean annual FLUXNET ET, water-balance-based ET (NLDAS-2 precipitation-USGS *Q*), and simulated ET from the four LSMs and their multimodel ensemble mean are compared for 12 RFCs in Figure 3. The results show that FLUXNET and water-balance-based ET (Figure 3f) is quite consistent among the 12 RFCs, with a very high correlation (0.998), small bias (−8.3 mm/year), and small RMSE (14.9 mm/year), suggesting that FLUXNET ET is a suitable ET reference data set for this study. The comparison clearly illustrates that Mosaic (Figure 3a) and SAC (Figure 3c) overestimate FLUXNET ET with large positive biases (130.5 mm/year for Mosaic and 89.9 mm/year for SAC). On the other hand, Noah (Figure 3b) and VIC (Figure 3d) underestimate the FLUXNET ET with large negative biases (−71.8 mm/year for Noah and −102.9 mm/year for VIC). The MME (Figure 3e) is closest to the FLUXNET ET with a bias of 11.4 mm/year, which is similar in magnitude to the water-balance-based ET (−8.4 mm/year), although there is a larger RMSE value (53.6 mm/year versus 14.9 mm/year).

When simulated *Q* from the four LSMs and their multimodel ensemble mean are compared with USGS *Q* (Figure 4), results opposite from those seen in the ET analysis are evident, as expected. Mosaic (Figure 4a)

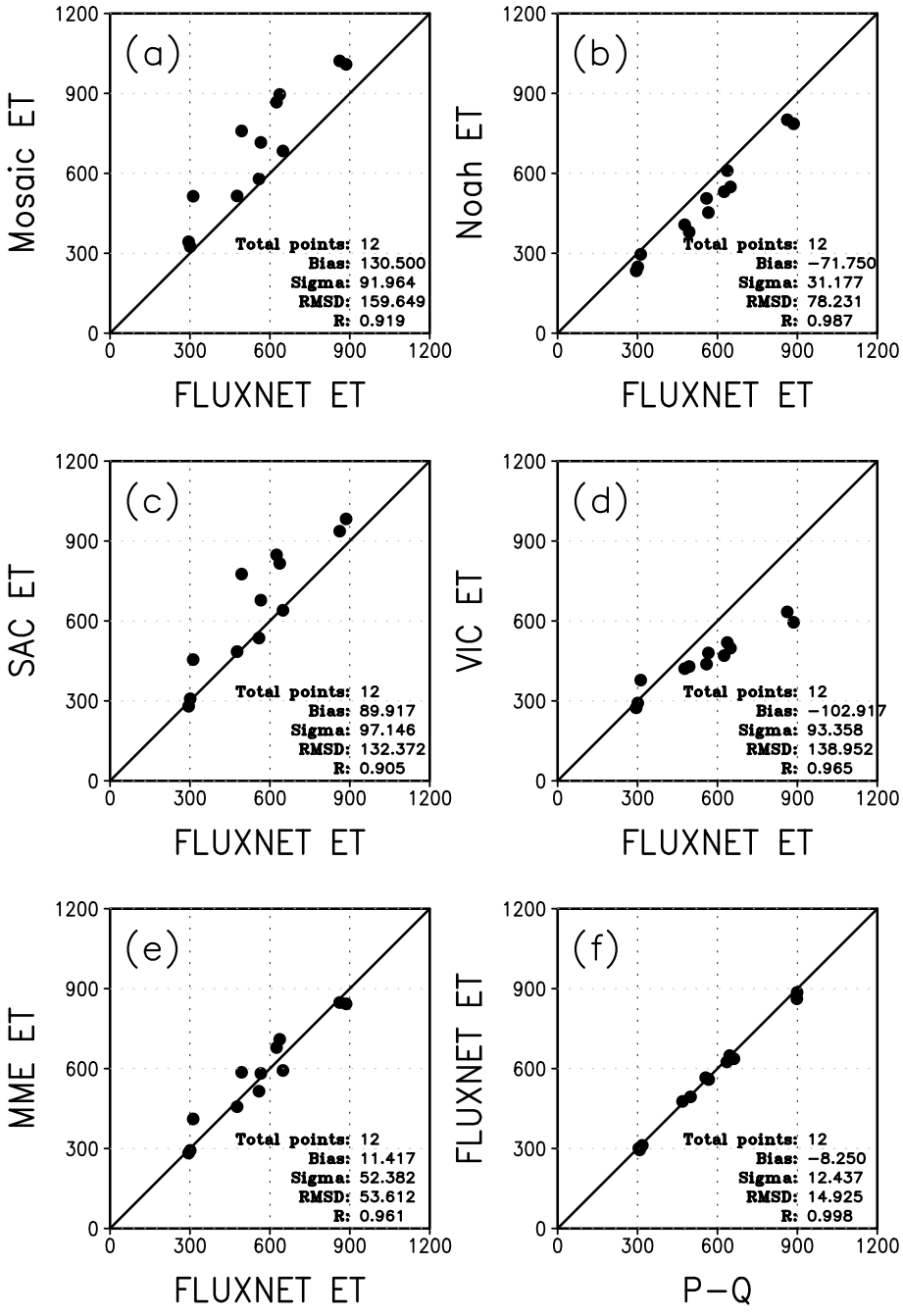

Figure 3. For each of 12 RFCs (dots), comparison of 27 year (1982–2008) mean annual ET (unit: mm/year) of MTE FLUXNET reference value (*x* axis) with that simulated by the operational NLDAS-2 LSMs of (a) Mosaic, (b) Noah, (c) SAC, (d) VIC, plus (e) their ensemble mean (MME), and (f) per the water budget as the difference of the observation-based NLDAS-2 *P* and USGS *Q*, with total water storage change (d*S*/d*t*) assumed negligible over 27 year annual period.

and SAC (Figure 4c) underestimate USGS *Q* with large negative biases (−130.3 mm/year for Mosaic and −80.5 mm/year for SAC). Noah (Figure 4b) and VIC (Figure 4d) overestimate USGS *Q* with large positive biases (71.8 mm/year for Noah and 111.0 mm/year for VIC). Again, the MME (Figure 4e) is closest to the USGS *Q* with a reasonable bias of −8.4 mm/year. The spatial distribution of the difference between mean annual observed FLUXNET ET (USGS *Q*) and the simulated ET (*Q*) of the four LSMs over all 12 RFC regions are shown in Figure 5 (left column) (Figure 5, right column). We recognize that there are some scale mismatch factors in these difference maps. However, we nevertheless can determine qualitatively which regions/RFCs have negative

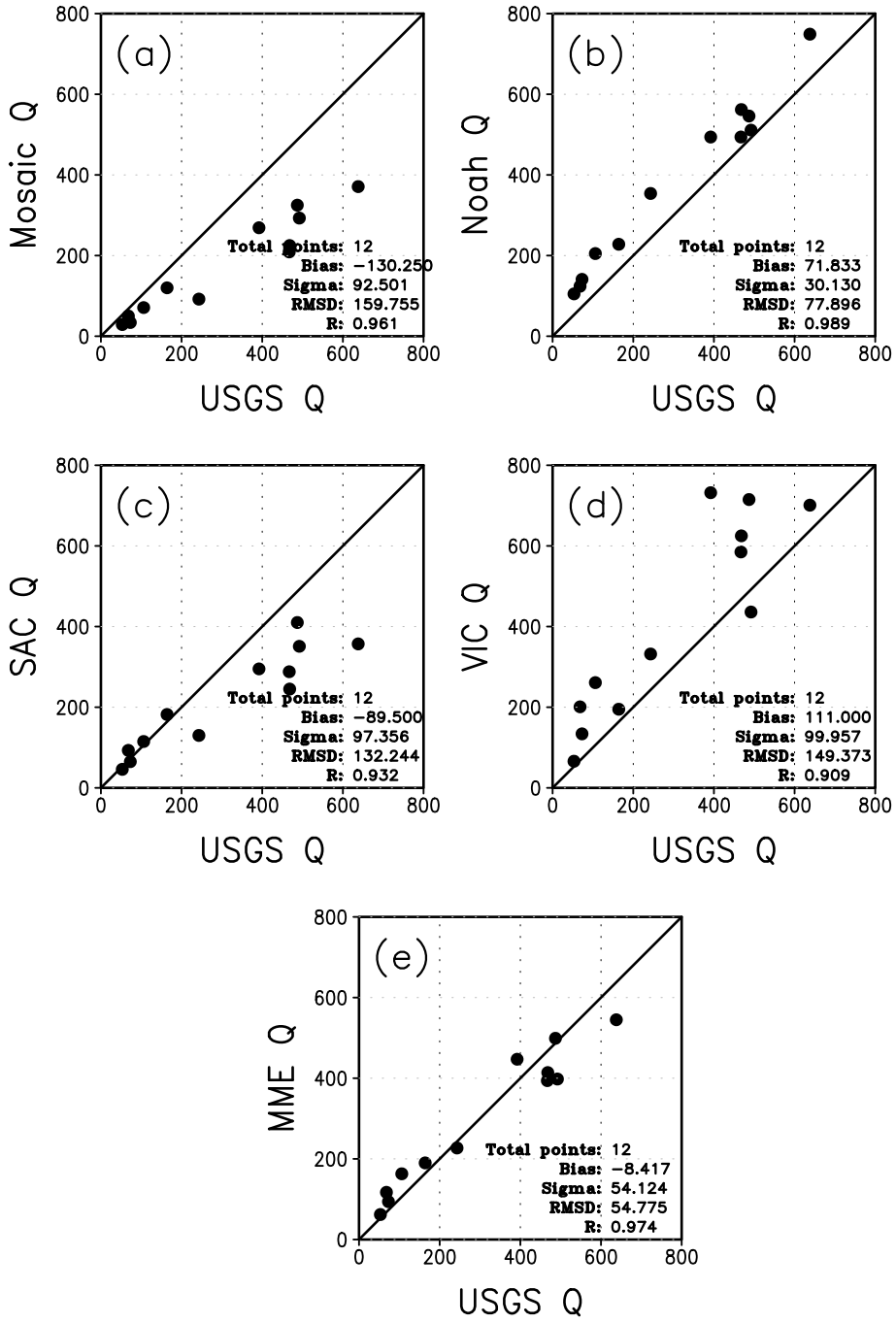

Figure 4. As in Figure 3 except for mean annual *Q* (unit: mm/year), with the observation-based USGS *Q* used as the reference value (*x* axis).

or positive biases. The results in Figure 5 are in good agreement with Figures 3 and 4 and provide additional detail regarding the spatial distribution of the LSM biases across the 12 RFCs.

The overall partitioning of mean annual precipitation into mean annual ET and *Q* for each LSM and the four-model MME for the 12 RFCs is shown in Figure 6. The partitioning can be validated over the 12 RFCs for which basin-observed total runoff is available. As discussed above, Noah and VIC overestimate the basin-observed total runoff at most of the 12 RFCs. By contrast, Mosaic and SAC underestimate the observations at the majority of the 12 RFCs. The MME is closest to these reference/observation products (Figure 6).

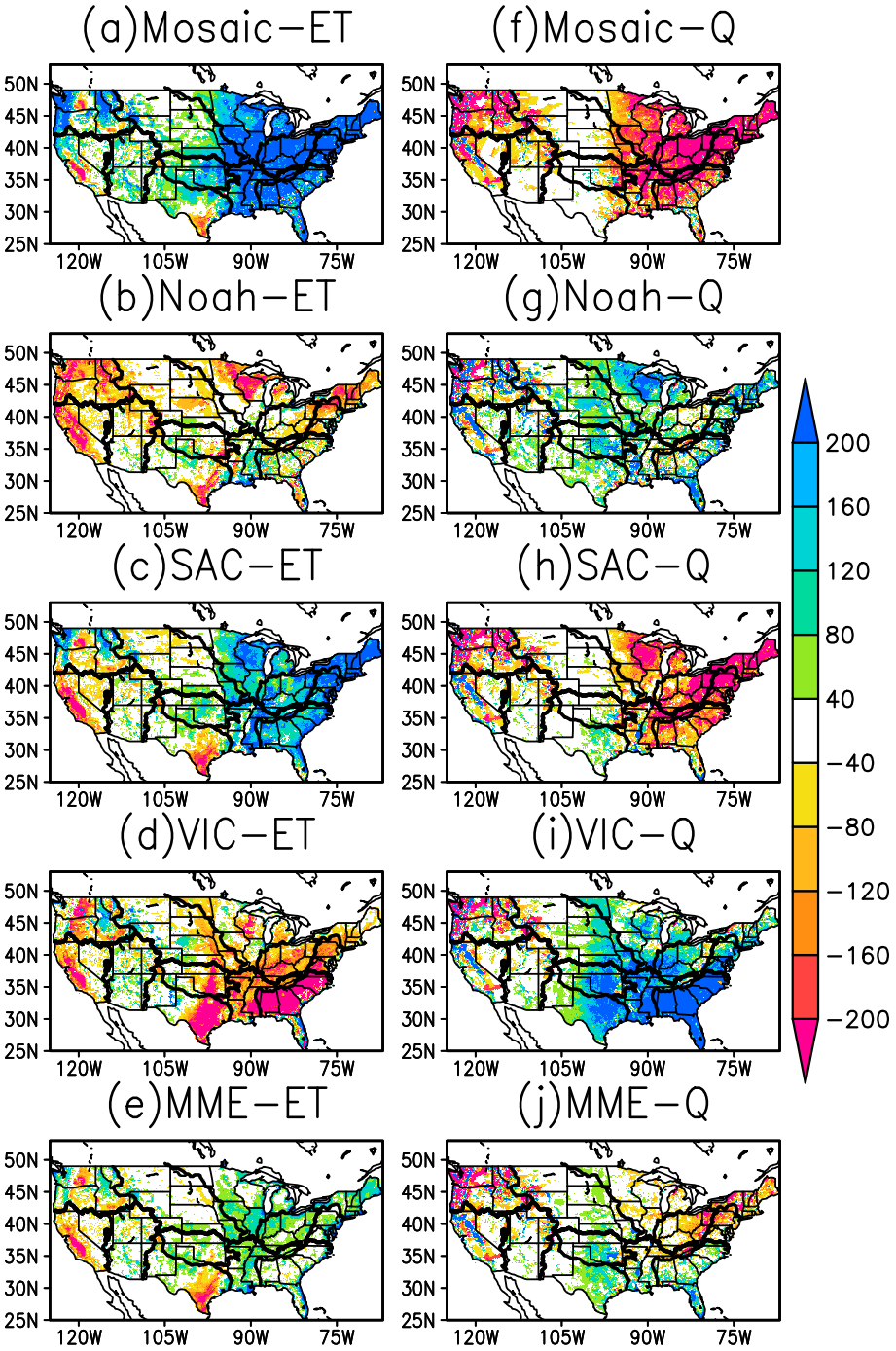

Figure 5. (left column) Difference between mean annual simulated ET of the operational NLDAS-2 from (a) Mosaic, (b) Noah, (c) SAC, (d) VIC, and (e) their ensemble mean (MME) and MTE FLUXNET ET. (right column) Difference between mean annual simulated *Q* of the operational NLDAS-2 from (f) Mosaic, (g) Noah, (h) SAC, (i) VIC, and (j) their ensemble mean (MME) and USGS *Q*. All results are for January 1982 to December 2008 (units: mm/year), as in Figure 2.

4.2. Analysis of Mean Monthly Annual Cycle

In studying the *monthly* water budget, we can no longer assume that the total water storage change term (dS/dt) is negligibly small as was assumed in the annual analysis. The term includes the storage change of total column soil moisture, snowpack, and canopy water in NLDAS-2 LSMs without considering the deep ground water storage change. While the deep ground water storage change is larger than the canopy water

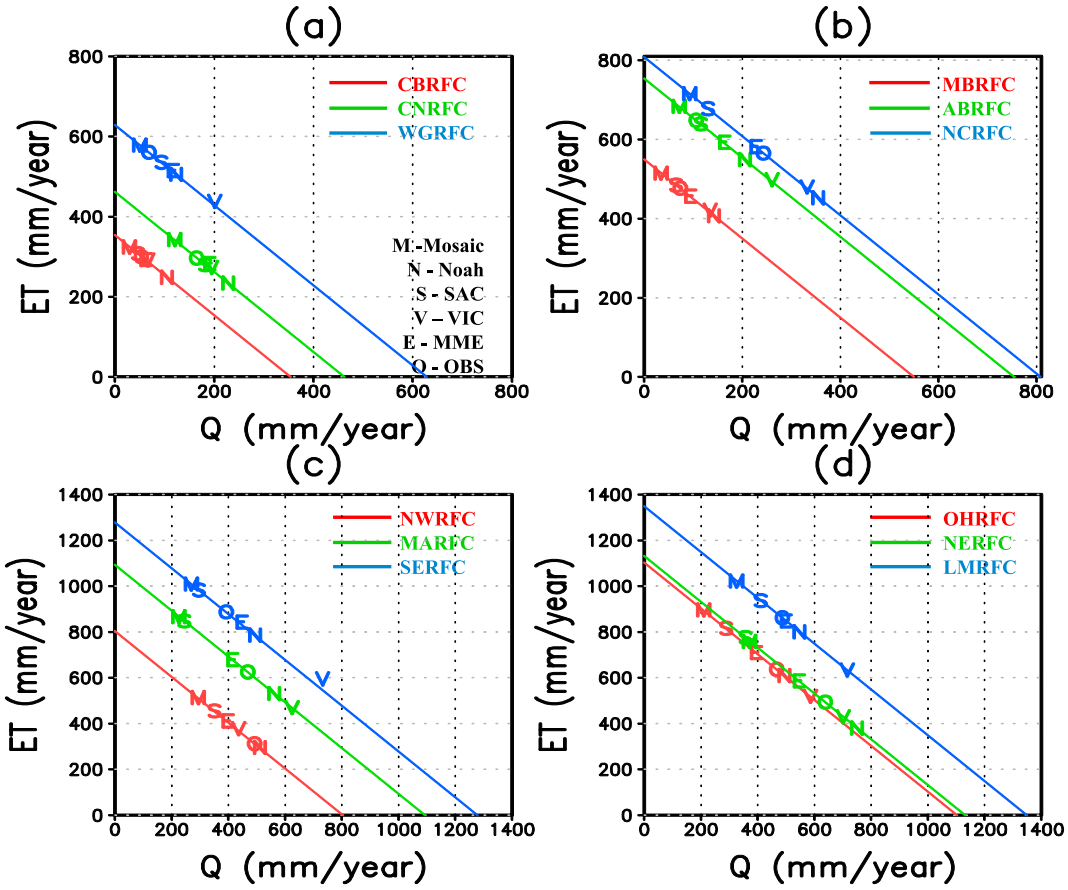

Figure 6. Partitioning of mean annual basin-mean precipitation (diagonal, mm/year) into mean annual basin-mean runoff *Q* (*x* axis) and evapotranspiration ET (*y* axis) for the 12 RFCs: (a) CBRFC, CNRFC, WGRFC, (b) MBRFC, ABRFC, NCRFC, (c) NWRFC, MARFC, SERFC, and (d) OHRFC, NERFC, LMRFC by Mosaic (M), Noah (N), SAC (S), VIC (V), MME (E), and observation-based USGS *Q* (O) for the period January 1982 to December 2008.

storage change, it is omitted from the equation since all four land surface models do not represent ground water processes. Even with this omission, the comparison of total water storage anomalies between the NLDAS-2 LSMs (e.g., Noah, VIC) and the Gravity Recovery and Climate Experiment (GRACE) [*Landerer and Swenson*, 2012] has shown good consistency [*Cai et al.*, 2014]. Forthcoming model improvements may address this shortcoming as representation of ground water processes is included in the Noah-MP model [*Niu et al.*, 2011; *Yang et al.*, 2011] and in the NASA/GSFC Catchment LSM (CLSM-F2.5) model [*Koster et al.*, 2000], which are being added to the NLDAS model suite.

Given the preceding discussion, when retaining the d*S*/d*t* term for analyzing the mean monthly annual water cycle, the land surface water budget equation can be expressed as

$$\mathrm{d}S/\mathrm{d}t = P - Q - \mathrm{ET} \tag{2}$$

where *P* is precipitation, *Q* is total runoff (surface and subsurface runoff), and ET is total evapotranspiration (from bare soil, canopy interception, and transpiration from vegetation). The reference total water storage d*S*/d*t* can be derived by using equation (2) for the 12 RFCs.

Figure 7 shows the mean annual cycle of FLUXNET ET and LSM-simulated ET from the operational NLDAS-2 system for the 12 RFCs. In general, Mosaic and SAC represent the upper bound and Noah and VIC represent the lower bound of the LSM ET envelope. Not surprisingly, the MME is closest to the FLUXNET ET. Mosaic largely overestimates the FLUXNET ET in the spring, early summer, and fall at almost all 12 RFCs, in particular at NWRFC, NCRFC, OHRFC, NERFC, and MARFC. Noah largely underestimates the FLUXNET ET in cold regions such as NWRFC, NCRFC, OHRFC, NERFC, MARFC, CNRFC, and CBRFC. SAC performs similarly to Mosaic, except for

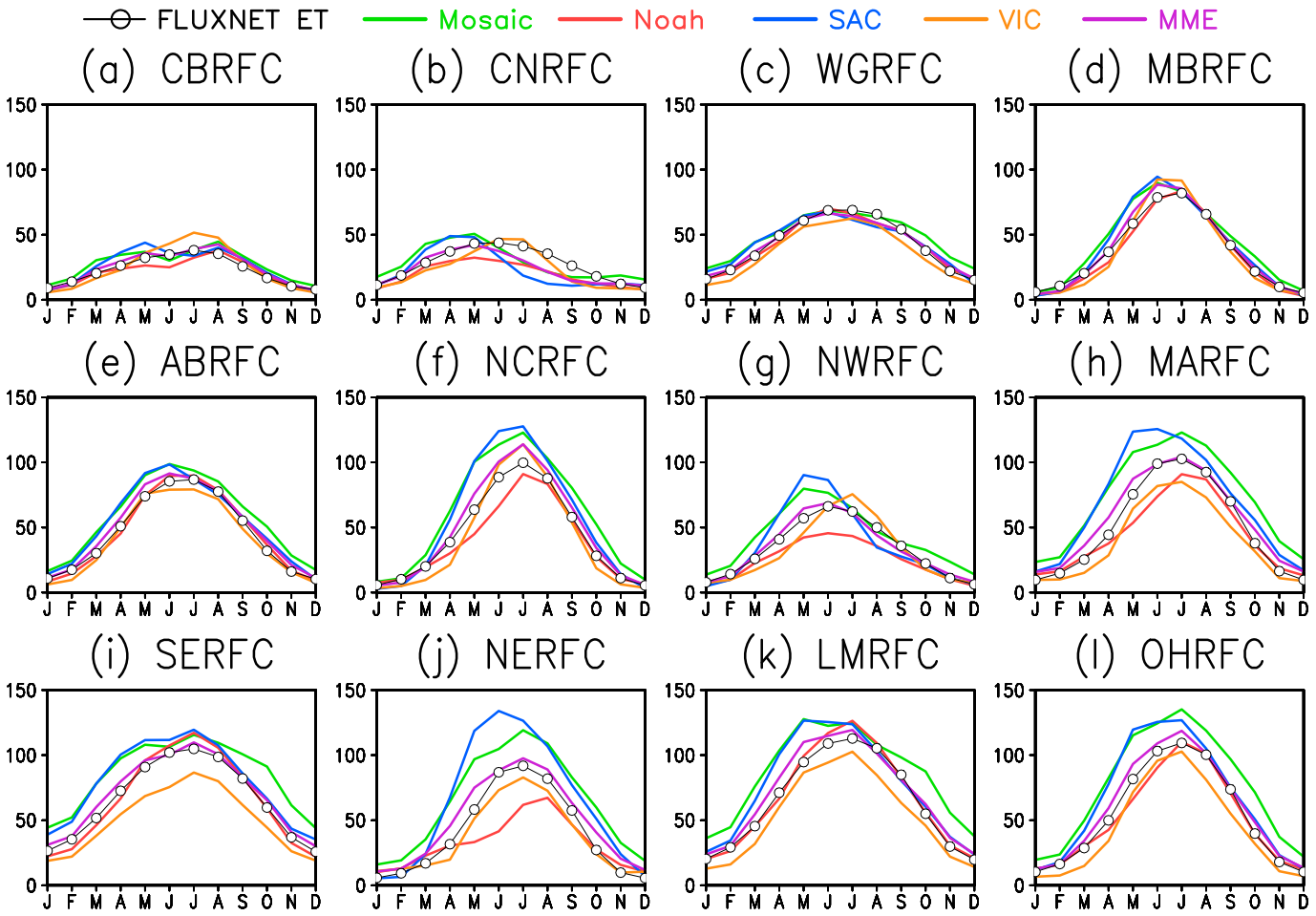

Figure 7. For each of the 12 RFCs, comparison of the 27 year (1982–2008) mean annual cycle of monthly mean ET (unit: mm/month) of the observation-based MTE FLUXNET reference (black line with open circles) with that simulated in the operational NLDAS-2 by the LSMs of Mosaic (green), Noah (red), SAC (blue), VIC (orange), and their ensemble mean MME (purple).

spring and early summer. VIC underestimates the FLUXNET ET in some RFCs over the eastern U.S. (i.e., OHRFC, MARFC, LMRFC, and SERFC). All four LSMs have difficulty capturing the seasonal cycle of FLUXNET ET at CNRFC and CBRFC, which are two mountainous RFCs and are furthermore the two driest RFCs according to the climate index listed in Table 3. At CNRFC, ET from Mosaic, Noah, and SAC seems to peak 1–3 months early, and VIC ET seems to peak 1 month late. The reason for the overly large LSM-simulated ET remains unclear. This issue needs to be investigated in the future. At CBRFC, the annual cycle of simulated ET in Noah and Mosaic substantially mimics the annual cycle of *P* (i.e., low ET and *P* in June and high ET and *P* in August), which is not reflected in the FLUXNET ET.

Figure 8 presents the mean annual cycle of 27 year monthly mean USGS *Q* and LSM-simulated and MME *Q* from the operational NLDAS-2 for the 12 RFCs. For *Q*, Mosaic and SAC give the lower bound while Noah and VIC give the upper bound of the model envelope of *Q*, which is the reverse of that for ET, as expected. In a broad sense, all four LSMs and their ensemble MME capture the annual cycle of monthly USGS *Q* for all the RFCs except CNRFC and CBRFC. However, the spread and disparity in *Q* across the four LSMs is very large,

Table 3. The 30 Year (1961–1990) Averaged Values of *P*/PET Climate Index (Adapted From *Schaake et al.* [2004])[a]

RFC	CB	CN	WG	MB	AB	NC	NW	MA	SE	NE	LM	OH
P/PET	0.29	0.37	0.37	0.50	0.54	0.82	0.96	1.03	1.04	1.22	1.29	1.33

[a]The RFCs are listed in order of increasing value of *P*/PET.

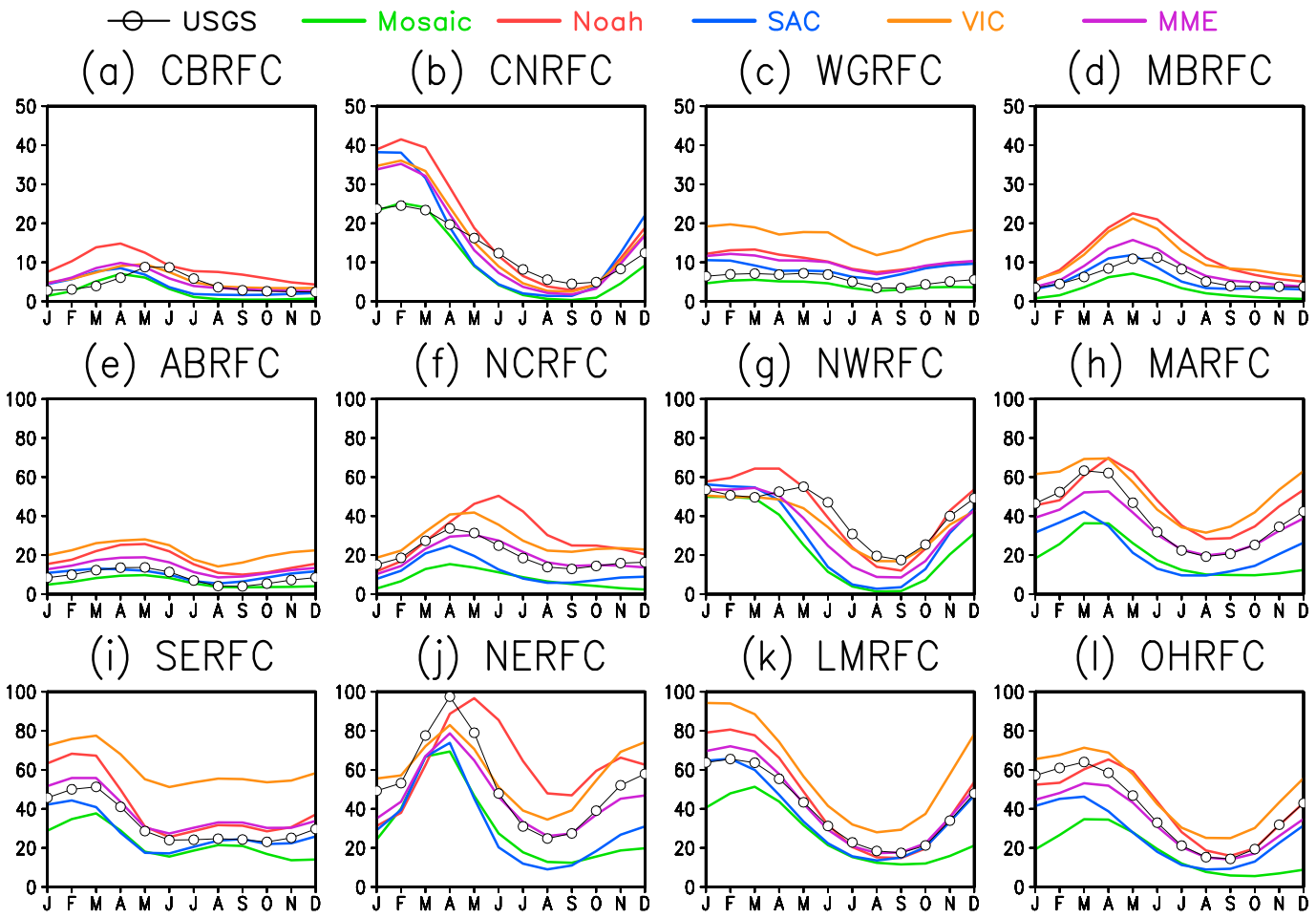

Figure 8. As in Figure 7 except for *Q*. The observation-based reference values (black line with open circles) are from USGS *Q*. The unit is mm/month.

which is consistent with the *Q* analysis of *Lohmann et al.* [2004]. The MME is quite close to the USGS *Q* and reasonably captures the USGS *Q* seasonal cycle for all 12 RFCs. The *Q* simulated by Noah peaks 1–2 months late when compared to USGS *Q* in the northeast quadrant of the CONUS (i.e., NCRFC, MARFC, OHRFC, and NERFC) and peaks 1–2 months early in MBRFC and CBRFC. This is likely due to the difficulty that the NLDAS-2 LSMs have in capturing the timing of snowmelt, as discussed in *Xia et al.* [2012b] and *Cai et al.* [2014], although the upgraded research versions have demonstrated improvements as shown in section 5. In addition, NLDAS-2 air temperature errors may give rise to an inaccurate partitioning of precipitation into rainfall and snowfall, as some studies have found that NLDAS-2 air temperature data have a 1–2°C warm bias when compared to in situ observations [*Royer and Poirier*, 2010]. The overly warm air temperatures lead to less snowfall and early snowmelt. However, the explanation for the delayed peak in *Q* which is seen over portions of the study area is less clear. Issues including forcing errors and model shortcomings need to be further investigated as part of future work.

Figure 9 presents the annual cycle of monthly mean LSM (and MME) simulated change in total water storage (dS/dt) compared to the water balance equation-derived dS/dt for each RFC. Generally speaking, all four LSMs reasonably capture the annual cycle of monthly mean dS/dt, although there are notable differences in the month of the peak simulated dS/dt across the four LSMs. VIC features better performance in simulating dS/dt than Mosaic, Noah, SAC, and the MME, which have similar performance. The Mosaic, Noah, SAC, and MME dS/dt peak 1 month earlier than the water balance equation-derived dS/dt.

4.3. Anomaly Correlation Analysis

Anomaly correlation (AC) is a useful means of representing the overall simulation skill of the four LSMs and the MME. The temporal anomaly is calculated from a 27 year monthly time series when the mean seasonal

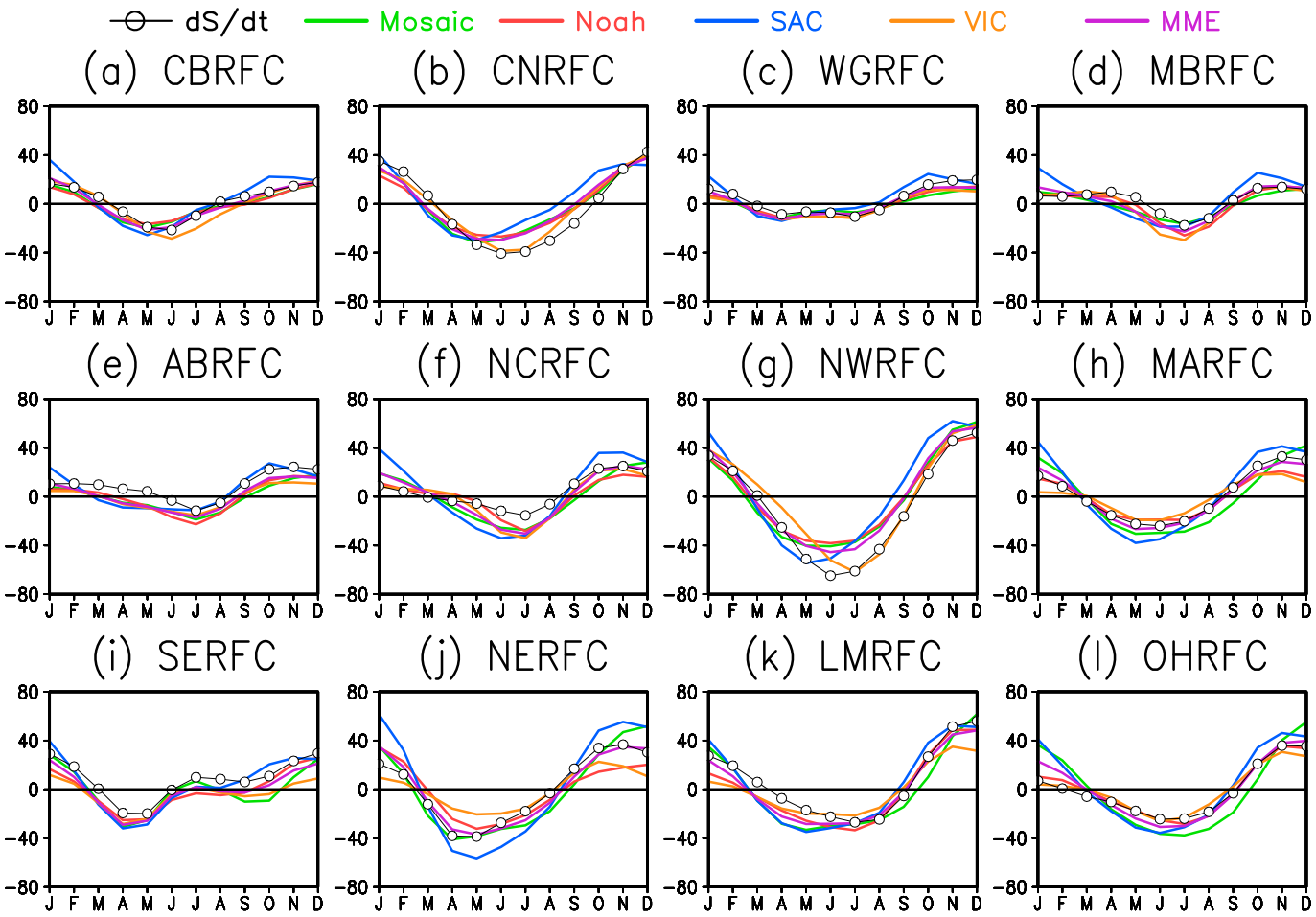

Figure 9. As in Figure 7 except for total water storage change (d*S*/d*t*, unit: mm/month). The observation-based reference values (black line with open circles) are calculated (per the surface water budget) as the difference between the NLDAS-2 precipitation and sum of the USGS reference *Q* and MTE FLUXNET reference ET.

cycle is removed. The temporal anomaly correlation is calculated from the observation-based anomaly (defined with respect to the observation-based climatology) and the simulated anomaly (defined with respect to the climatology of a given LSM's simulation). Table 4 lists the AC values for *Q*, ET, and d*S*/d*t* for each RFC. Using the USGS *Q* as the reference for computing the temporal correlation of the AC values for simulated *Q*, the MME has the largest AC and hence the best performance at 9 of 12 RFCs, followed by VIC, Noah, SAC, and Mosaic. All models display their worst AC performance in *Q* at CBRFC and their best performance at LMRFC. One reason for the lower LSM AC values of *Q* for CBRFC is the poor LSM simulation of the annual cycle of monthly mean *Q* (Figure 8h). The AC values for d*S*/d*t* are quite high for almost all basins and models (>0.90) except for VIC at LMRFC (0.89) and SERFC (0.88), suggesting that all models are skillful at capturing the anomalies of total water storage change. The relatively low AC values for VIC at LMRFC and SERFC are due mainly to suboptimal VIC conceptual soil and hydrology parameters, which include the variable infiltration curve parameter, maximum baseflow velocity (Dsmax), fraction of Dsmax where nonlinear baseflow begins, fraction of maximum soil moisture content above which nonlinear baseflow occurs, layer 2 and layer 3 soil depth, and the parameter characterizing the variation of saturated hydraulic conductivity with soil moisture. Using calibrated parameters [*Troy et al.*, 2008] yields increases in AC values from 0.89 to 0.95 for LMRFC and from 0.88 to 0.96 for SERFC (see Table 7).

For the analysis of the AC of simulated ET with respect to the FLUXNET ET, all LSMs show a lower AC for ET over all 12 RFCs (except for CBRFC) than for the AC for *Q*, in particular at NCRFC, OHRFC, NERFC, MARFC, LMRFC, and SERFC. A likely reason for this lower performance of simulated ET relative to *Q* is the weak correlation between precipitation and ET anomalies in those regions [*Xia et al.*, 2012c]. The relatively high AC values

Table 4. Anomaly Correlation (AC) for Runoff *Q* (Top Section), Evapotranspiration ET (Middle Section), and Total Water Storage Change d*S*/d*t* (Bottom Section) Between Observed and Modeled Water Budget Components in the NCEP Operational NLDAS-2 for the 27 Year Period of 1982 to 2008[a]

RFC	CBRFC	CNRFC	WGRFC	MBRFC	ABRFC	NCRFC	NWRFC	MARFC	SERFC	NERFC	LMRFC	OHRFC
						Q						
Mosaic	0.59	0.94	0.87	0.83	0.85	0.81	0.87	0.86	0.94	0.85	0.90	0.79
Noah	0.68	**0.95**	0.87	0.89	0.91	0.87	0.92	0.91	0.95	0.83	**0.97**	0.87
SAC	0.57	0.89	0.88	0.86	0.92	0.90	0.89	0.91	0.91	0.83	0.93	0.90
VIC	**0.76**	**0.95**	**0.90**	0.85	0.92	0.94	**0.93**	0.92	0.93	**0.94**	**0.97**	**0.95**
MME	0.69	0.94	**0.90**	**0.90**	**0.94**	**0.95**	**0.93**	**0.95**	**0.96**	**0.94**	**0.97**	0.94
						ET						
Mosaic	**0.83**	0.80	0.85	0.74	0.77	0.42	0.58	0.29	0.04	0.36	0.16	0.41
Noah	0.81	**0.81**	0.87	0.79	**0.81**	0.48	0.60	**0.41**	**0.36**	0.21	**0.36**	**0.52**
SAC	0.72	0.69	0.85	0.79	0.76	0.21	0.58	0.06	0.03	0.10	0.05	0.09
VIC	0.75	0.65	0.83	0.82	0.73	**0.59**	0.59	0.15	0.16	**0.48**	0.20	0.38
MME	0.82	0.80	**0.88**	**0.84**	**0.81**	0.49	**0.67**	0.22	0.12	0.36	0.19	0.39
						d*S*/d*t*						
Mosaic	0.90	**0.97**	0.92	0.93	0.95	0.93	0.93	0.93	**0.96**	0.91	0.94	0.86
Noah	0.93	0.95	**0.96**	**0.97**	**0.97**	0.94	0.94	0.92	0.94	0.86	0.96	0.91
SAC	0.90	0.93	0.91	0.93	0.95	0.94	0.93	0.93	0.95	0.93	0.96	0.92
VIC	**0.96**	0.94	0.94	0.93	0.95	**0.96**	**0.96**	0.89	0.88	0.94	0.93	0.93
MME	0.94	0.96	0.94	0.96	**0.97**	**0.96**	**0.96**	**0.95**	**0.96**	**0.96**	**0.97**	**0.94**

[a]The value in bold font in each column of each section denotes the maximum value for the given RFC (an AC value $>$ |0.12| is significant at the 5% significance level).

for ET, which is evident for MBRFC, CNRFC, CBRFC, ABRFC, and WGRFC, are due to higher correlations between precipitation and ET in those regions [*Xia et al*., 2012c].

These instances of relatively higher (lower) AC for ET are associated with dry (wet) RFCs with desert (forest) characteristics. Table 3 lists the climatological aridity for the 1961–1990 period for each RFC [*Dooge*, 1997; *Sankarasubramanian and Vogel*, 2002]. An index value below 0.5 indicates a desert climate, while a value between 0.5 and 1.0 indicates a grassland ecosystem, and a value above 1.0 indicates a forest ecosystem. Five of the six RFCs with low AC for simulated ET have an aridity index characteristic of wetter/forested regions. All five RFCs with relatively higher AC for simulated ET have aridity index values indicative of semiarid or desert regions. For wetter/forested RFCs, ET variations are mainly determined by air temperature, wind speed, relative humidity, and incoming solar radiation (rather than mainly by precipitation as in semiarid/desert RFCs); and therefore ET variations are more difficult to simulate for wetter/forested RFCs and remain a challenge for the land surface modeling community.

As stated at the 2015 American Meteorological Society Annual Meeting Horton Lecture (C. Milly, personal communication, 2015), the use of different algorithms for calculating potential evapotranspiration (PET) can lead to extremely disparate PET estimates. Such disparate PET values would contribute strongly to wide disparities in simulated actual ET across a collection of LSMs. This issue of the choice of algorithm to calculate PET may partially explain why, as detailed below, SAC-Clim performed better than the SAC version used in the operational NLDAS-2, since the operational NLDAS-2 SAC uses Noah-derived PET rather than the observation-based climatological PET.

4.4. Comparison of Monthly GRACE Observed and NLDAS-2 Simulated TWSA and TWSC

In this section, we compare and analyze the 2003–2014 monthly GRACE-observed and model-simulated TWSA for the four-model ensemble mean at 12 RFCs (Figure 10). The results show that the MME captures the monthly variation of GRACE-observed TWSA quite well. In particular, the MME captures large depletions of total water storage at CNRFC, WGRFC, and ABRFC, which is consistent with the previous study in Texas [*Long et al*., 2013]. At almost all of the RFCs, the amplitude of the simulated TWSA is smaller than the GRACE-observed TWSA. This is true in particular at wet RFCs. This is likely due to the fact that the models do not represent the impact of groundwater on TWSA. At LMRFC, the MME exhibits much smaller amplitude than the GRACE-observed TWSA. Besides lacking representation of groundwater, the models do not explicitly model the change of water storage in rivers/lakes and the exchange of groundwater and river/lake water. This becomes an issue along the Mississippi River, which is characterized by very high amplitude variations

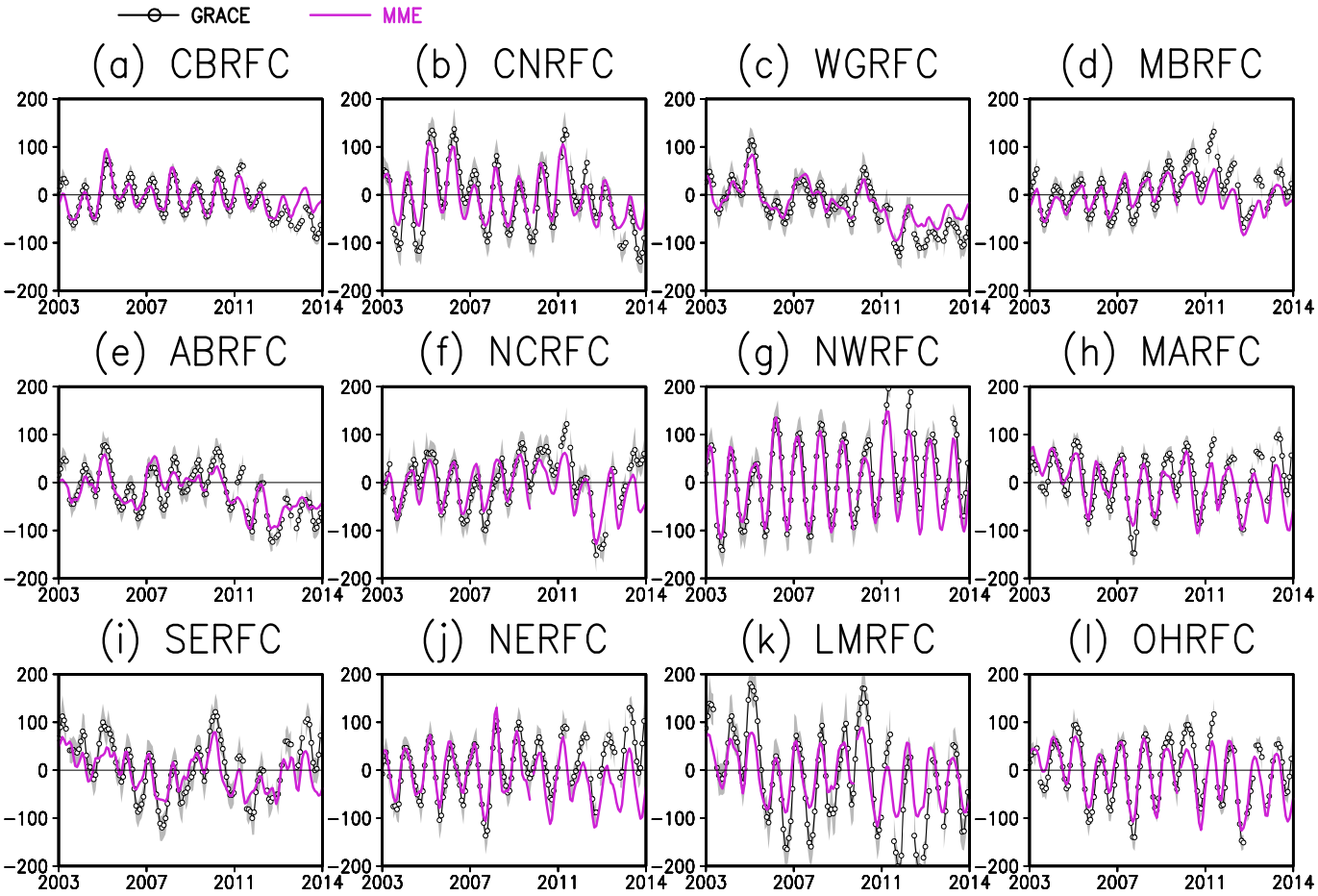

Figure 10. For each of the 12 RFCs, comparison of the 12 year (2003–2014) time series of monthly total water storage anomaly (TWSA, unit: mm) from the GRACE-derived data set (thin black line with open circles) and that simulated by the multimodel ensemble MME (purple solid line) of the operational NLDAS-2. The grey shaded area represents one standard deviation of the GRACE-derived value, as an indicator of the uncertainty of the GRACE TWSA values. The NLDAS-2 climatology underpinning the simulated TWSA values is calculated from January 2004 to December 2009, matching the period used by GRACE Tellus. We note that the MME-simulated total water storage includes only total column soil moisture, snow water equivalent, and canopy water storage, thus omitting ground water and reservoir storage. The GRACE total water storage includes total column soil moisture, snow water equivalent, canopy water storage, ice, reservoir storage (e.g., rivers, lakes, and ponds), and ground water.

Table 5. Anomaly Correlation (AC) Coefficient Between GRACE-Observed and NLDAS-2 Simulated Total Water Storage Anomaly (TWSA) Is Calculated for 12 RFCs From January 2003 to December 2014[a]

RFC Name	Mosaic	Noah	SAC	VIC	MME
CBRFC	0.78	0.75	0.77	**0.83**	0.80
CNRFC	0.89	0.89	0.81	**0.90**	0.89
WGRFC	0.88	**0.92**	0.87	0.86	0.90
MBRFC	0.82	**0.86**	0.66	0.84	0.83
ABRFC	0.83	**0.90**	0.80	0.81	0.86
NCRFC	0.83	**0.83**	0.56	0.75	0.77
NWRFC	0.89	0.91	0.81	**0.96**	0.92
MARFC	0.77	**0.79**	0.60	0.65	0.74
SERFC	0.79	**0.83**	0.53	0.62	0.77
NERFC	0.74	**0.76**	0.63	0.68	0.73
LMRFC	**0.90**	0.86	0.78	0.79	0.88
OHRFC	0.87	**0.88**	0.75	0.83	0.87
Mean	0.83	0.85	0.71	0.79	0.83

[a]The bold font denotes the maximum AC values from Mosaic, Noah, SAC, VIC, and MME for each RFC (an AC value $> |0.12|$ is significant at the 5% significance level).

in total water storage due to water level fluctuations [*Cai et al.*, 2014]. The AC values between GRACE-observed and model-simulated TWSA are listed for each individual LSM in Table 5. The results show that Noah performs the best of the NLDAS-2 models as it exhibits the largest AC value at eight of the RFCs. The largest factor in this performance is the result of a reasonable simulation of soil moisture [*Xia et al.*, 2014a, 2015a] as (1) canopy water storage plays very small role and (2) simulation of SWE in Noah is comparable with the other three models [*Xia et al.*, 2012a]. VIC features the best performance at CBRFC, CNRFC, and NWRFC. Here VIC exhibits a superior

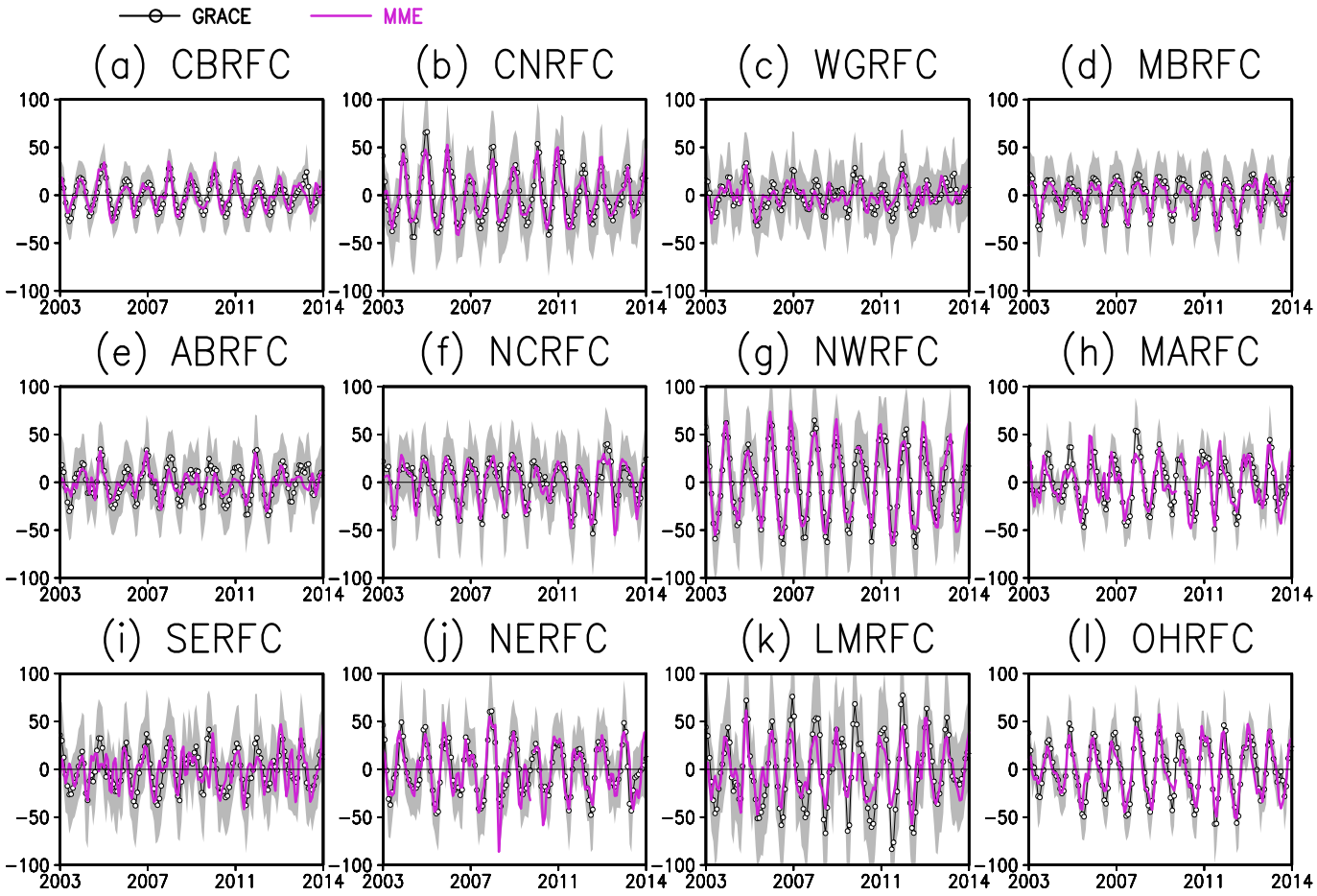

Figure 11. As in Figure 10 but for total water storage change (TWSC, units: mm/month).

simulation of SWE due to its inclusion of more advanced snowpack processes (e.g., snow bands, multiple snow layers; see *Xia et al*. [2012a, 2012b]). The Mosaic LSM features the largest AC value at LMRFC, indicating that it has the best performance at that basin. The mean results show that the Noah model has the largest AC value (0.85), followed by the Mosaic and MME (0.83), VIC (0.79), and SAC (0.71).

Table 6. Anomaly Correlation (AC) Coefficients Between GRACE-Observed and NLDAS-2 Simulated Total Water Storage Change (TWSC) Are Calculated at 12 RFCs From January 2003 to December 2014[a]

RFC Name	Mosaic	Noah	SAC	VIC	MME
CBRFC	0.64	0.62	0.61	**0.78**	0.69
CNRFC	0.78	0.77	0.73	**0.82**	0.79
WGRFC	0.49	**0.57**	0.47	0.49	0.50
MBRFC	0.70	**0.77**	0.71	0.76	0.76
ABRFC	0.56	**0.65**	0.46	0.47	0.55
NCRFC	0.69	0.69	0.69	0.69	**0.71**
NWRFC	0.84	0.86	0.83	**0.91**	0.88
MARFC	**0.65**	0.47	0.61	0.35	0.56
SERFC	0.36	**0.49**	0.41	0.27	0.39
NERFC	0.71	0.61	**0.73**	0.47	0.68
LMRFC	**0.72**	0.70	0.67	0.55	0.68
OHRFC	**0.78**	0.67	0.74	0.60	0.73
Mean	0.66	0.66	0.64	0.60	0.66

[a]The bold fonts represent maximum anomaly correlation values from Mosaic, Noah, SAC, VIC, and MME for each RFC (an AC value $> |0.12|$ is significant at the 5% significance level).

Model-simulated TWSC was also evaluated against GRACE-derived TWSC (Figure 11 and Table 6). The simulated TWSC is calculated from equation (2) using monthly precipitation, total runoff, and evapotranspiration. GRACE-derived TWSC appears to be noisier and have lower amplitude than GRACE-observed TWSA because the latter is the primitive integral of TWSC (see Figures 10 and 11). The magnitude of the uncertainty in GRACE-derived TWSC is amplified compared with that in the GRACE-observed TWSA, although a centered difference derivative method is used

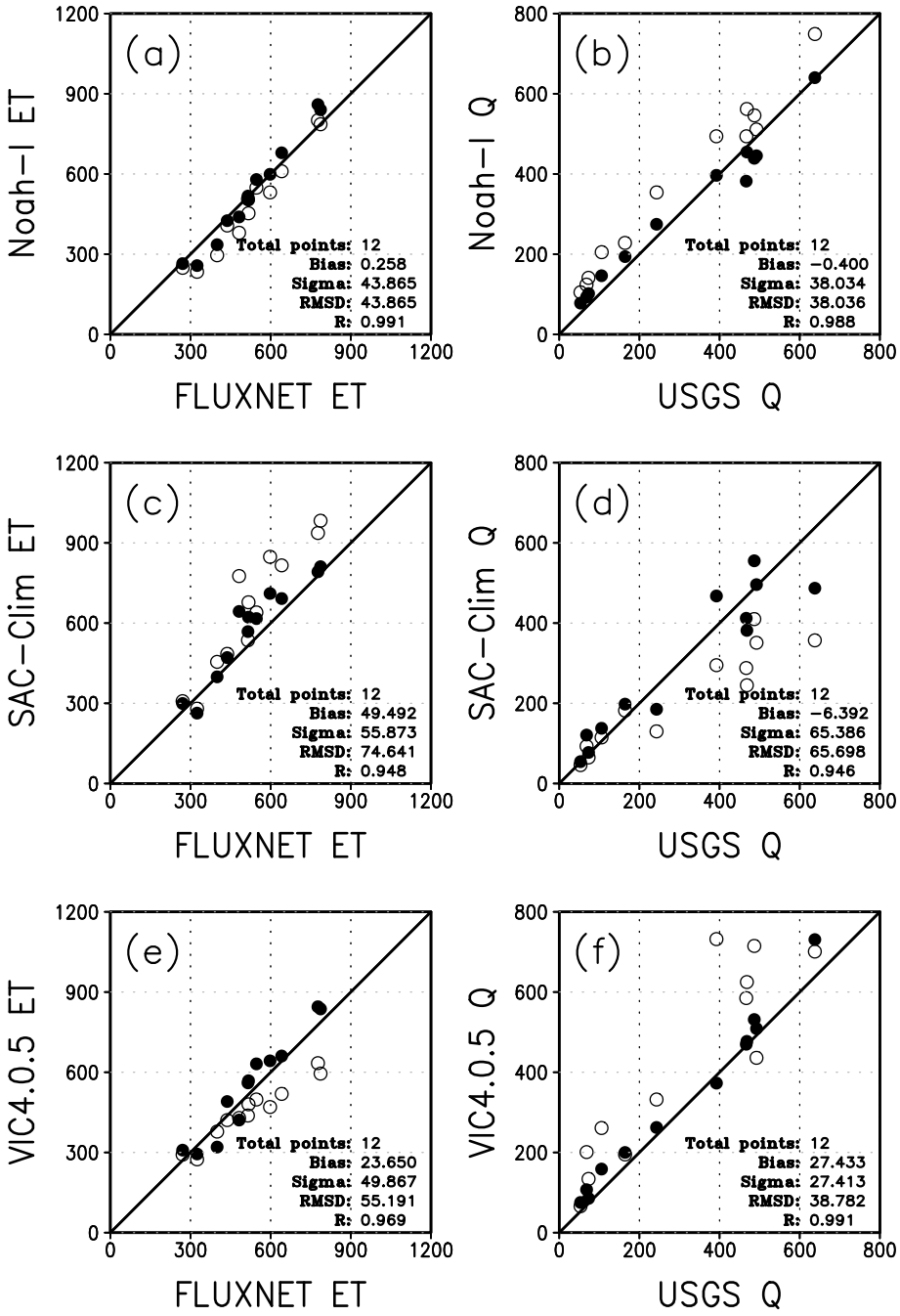

Figure 12. (left column) For each of 12 RFCs, comparison of 27 year (1982–2008) mean annual ET (unit: mm/year) of MTE FLUXNET reference value (*x* axis) with that simulated in the NLDAS-2 by the operational LSMs (open circles) and the research LSMs (closed circles) of Noah and Noah-I (top), SAC and SAC-Clim (middle), and VIC and VIC4.0.5 (bottom). (right column) As in Figure 12 (left column) except for *Q*, with the observation-based USGS *Q* as the reference value. The given statistical metrics are for the research LSMs. Figures 3 and 4 give the metrics for the operational LSMs.

[*Long et al.*, 2014]. The results show that the MME is able to capture the broad features and monthly variability of the GRACE-derived TWSC over the 12 year study period. Compared with the results of the TWSA evaluation, all models exhibit smaller correlation coefficients across all of the RFCs (Table 6). The maximum correlation values are dispersed more widely across LSMs than are the TWSA AC analysis (MARFC, LMRFC, and OHRFC for Mosaic; WGRFC, MBRFC, ABRFC, and SERFC for Noah; NERFC for SAC; CBRFC, CNRFC, and NWRFC for VIC;

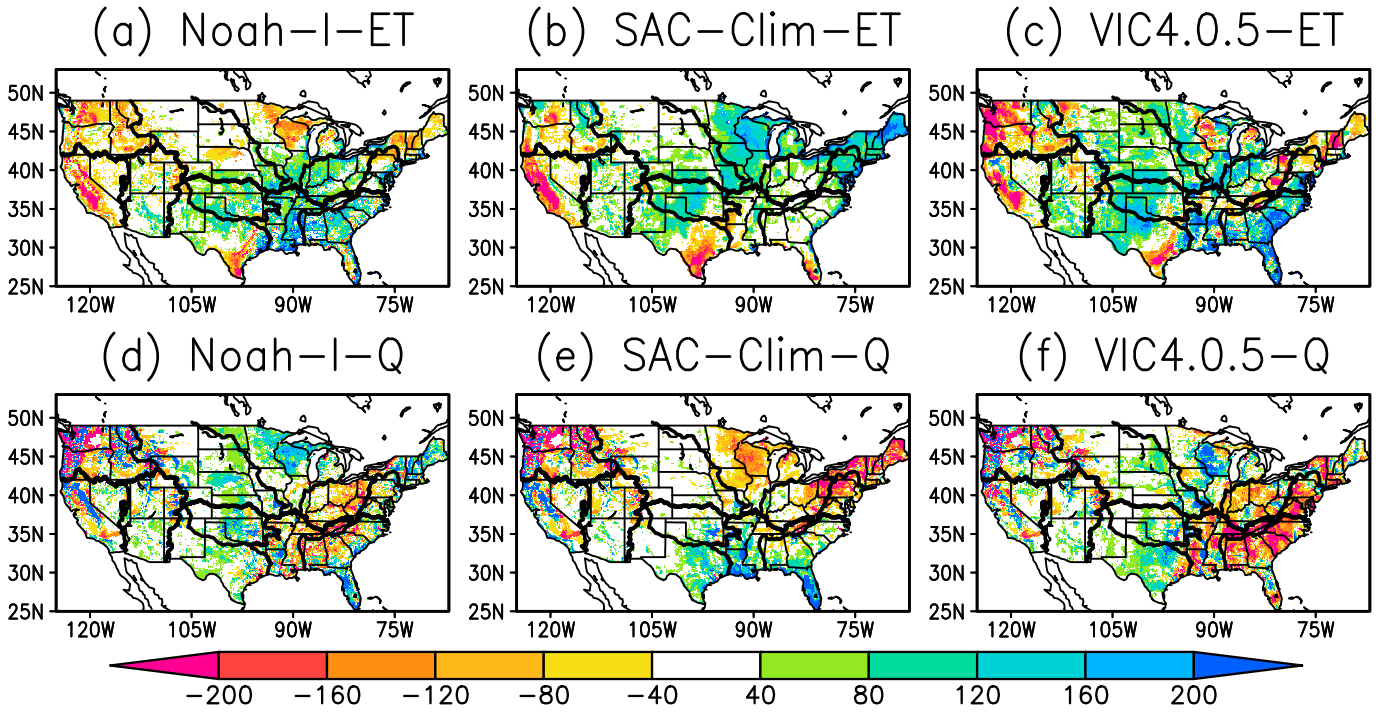

Figure 13. (top row) Difference between mean annual ET (mm/year) simulated in the research NLDAS-2 by LSMs of (a) Noah-I, (b) SAC-Clim, and (c) VIC4.0.5 and the observation-based MTE FLUXNET reference ET. (bottom row) Difference between mean annual *Q* (mm/year) simulated in the research NLDAS-2 by LSMs of (d) Noah-I, (e) SAC-Clim, and (f) VIC4.0.5 and observation-based USGS reference *Q*.

and NCRFC for MME). The VIC model features very consistent results for the three western RFCs, suggesting that the VIC model exhibits strong skill in simulating both TWSA and TWSC at these locations. The overall performance for all of the models is very similar although SAC and VIC have slightly lower average correlation coefficients (see Table 6).

5. Evaluation of Water Budget Components for Research NLDAS-2

5.1. Mean Annual Climatology Analysis

For the research NLDAS-2 system, Noah, SAC, and VIC have been upgraded to Noah-I, SAC-Clim, and VIC4.0.5, respectively. The purpose of this section is to examine whether these upgrades can reduce the biases in LSM-simulated mean annual ET and *Q*. The mean annual simulated ET and *Q* from the three upgraded models are compared to mean annual FLUXNET ET and USGS *Q* for the 12 RFCs in Figure 12. For ET, the bias (RMSE) is very substantially reduced from −71.8 mm/year to 0.3 mm/year (from 78.2 mm/year to 43.6 mm/year) for Noah-I, is significantly reduced from 89.9 mm/year to 49.5 mm/year (from 132.4 mm/year to 74.6 mm/year) for SAC-Clim, and is greatly reduced from −102.9 mm/year to 23.7 mm/year (from 139.0 mm/year to 55.2 mm/year) for VIC4.0.5. Thus, the RMSE is reduced by more than 50% for all three upgraded models, suggesting a huge improvement. In the *Q* comparison in Figure 12, a similar reduction is also evident. The spatial distribution of the bias of simulated ET and *Q* in the three upgraded LSMs is shown for the 12 RFCs in Figure 13. In comparison to Figure 5, large negative (positive) ET (*Q*) biases in mountainous RFCs having substantial cold-season snowpack are largely reduced for Noah-I. For SAC-Clim, large positive (negative) ET (*Q*) biases have been reduced to moderate negative (positive) biases in the eastern U.S. For VIC4.0.5, large negative (positive) ET (*Q*) biases have been reduced to moderate positive (negative) ET (*Q*) biases in the southeastern U.S., as the calibrated parameters [*Troy et al.*, 2008] used in the study affect not only the simulation of total runoff but also of soil moisture and ET. This leads to a negative-to-positive change in the ET bias in the southeast. However, the ET and *Q* performance of all three upgraded models over the mountainous Sierra Nevada region are less improved, as large biases still exist there. Further efforts to improve LSMs and forcing data are thus called for over that region.

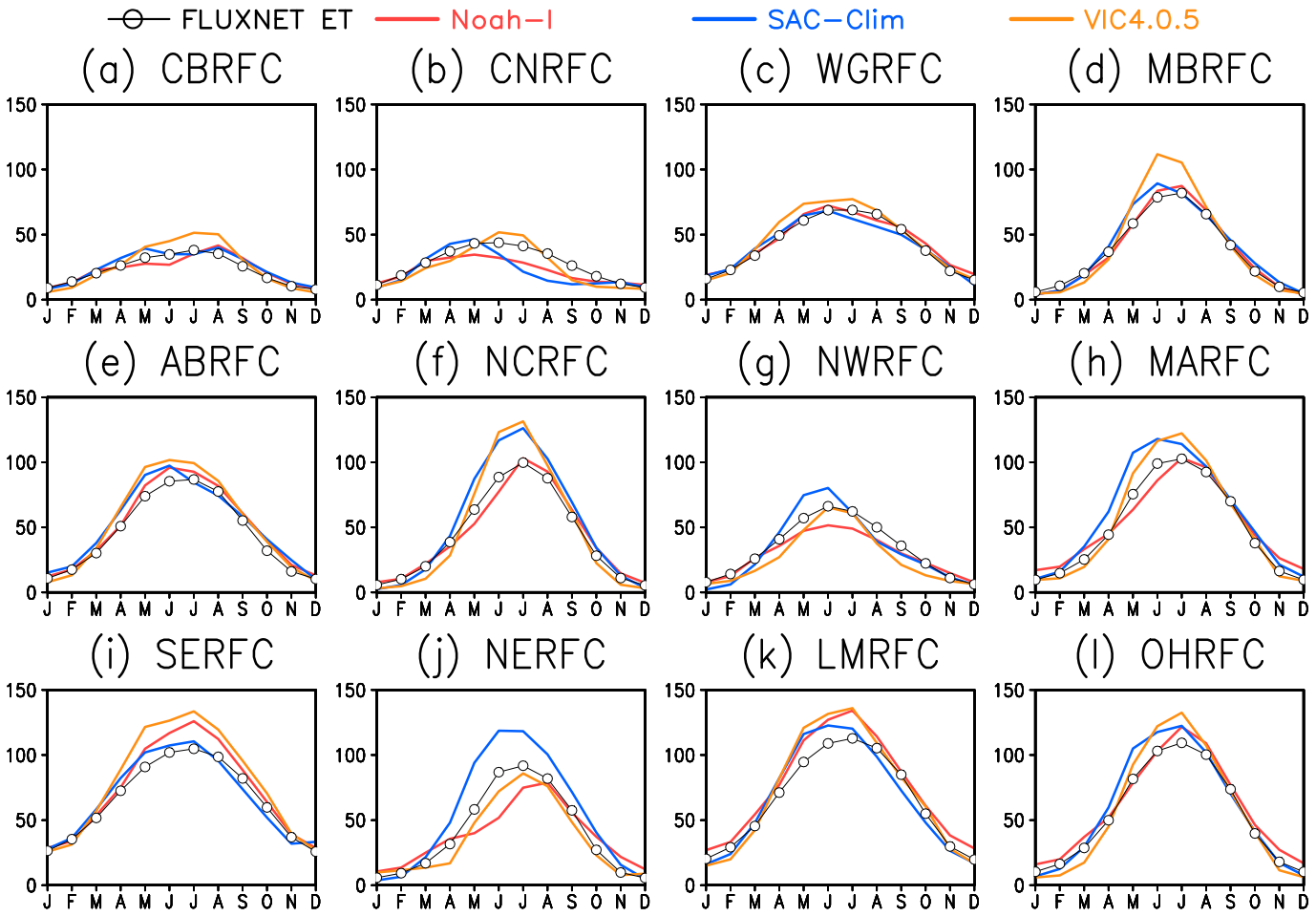

Figure 14. For each of the 12 RFCs, comparison of the 27 year (1982–2008) mean annual cycle of monthly mean ET (unit: mm/month) of the observation-based MTE FLUXNET reference (black line with open circles) with that simulated in the research NLDAS-2 by the LSMs of Noah-I (red), SAC-Clim (blue), and VIC4.0.5 (orange).

5.2. Mean Annual Cycle Analysis

In Figure 14, the annual cycle of mean monthly ET simulated by the three upgraded LSMs is compared with that of the observation-based FLUXNET ET. As expected, the simulated ET is closer to the FLUXNET ET in the three upgraded LSMs when compared to Noah, SAC, and VIC output from the operational NLDAS-2, and there is a smaller disparity or spread. In spite of these improvements, Noah-I still underestimates the FLUXNET ET at NWRFC, NERFC, CNRFC, and CBRFC in the warm season (May–September). We propose that this should be addressed through further Noah LSM community model development, given the longtime challenges that the EMC Land Team and its collaborators have faced to solve this Noah ET bias over the four cited RFCs. SAC-Clim still overestimates the FLUXNET ET at NWRFC, NCRFC, NERFC, and MARFC, although this overestimation is greatly reduced when compared to SAC in the operational NLDAS-2 (Figure 7). The VIC4.0.5 ET simulation is largely improved, in particular over MARFC, LMRFC, and SERFC when compared to the VIC ET simulation in the operational NLDAS-2. As is the case for simulated ET, the *Q* simulation from Noah-I, SAC-Clim, and VIC4.0.5 is also largely improved when compared to those in the operational NLDAS-2 system, although there are still some inaccuracies in the simulated versus observed month of peak *Q* (Figure 15).

5.3. Anomaly Correlation and Nash-Sutcliffe Efficiency Analysis

Table 7 presents the AC values of simulated monthly mean ET, *Q*, and d*S*/d*t* in the three upgraded LSMs as computed with respect to the 27 year time series of observed monthly anomalies obtained from the observation-based monthly mean ET, *Q*, and d*S*/d*t* reference data sets. For the AC of monthly mean ET simulated by the three upgraded LSMs, six (three) RFCs show significant improvements (deteriorations) for Noah-I, two (three) RFCs show significant improvements (deteriorations) for SAC-Clim, and three (four) RFCs show

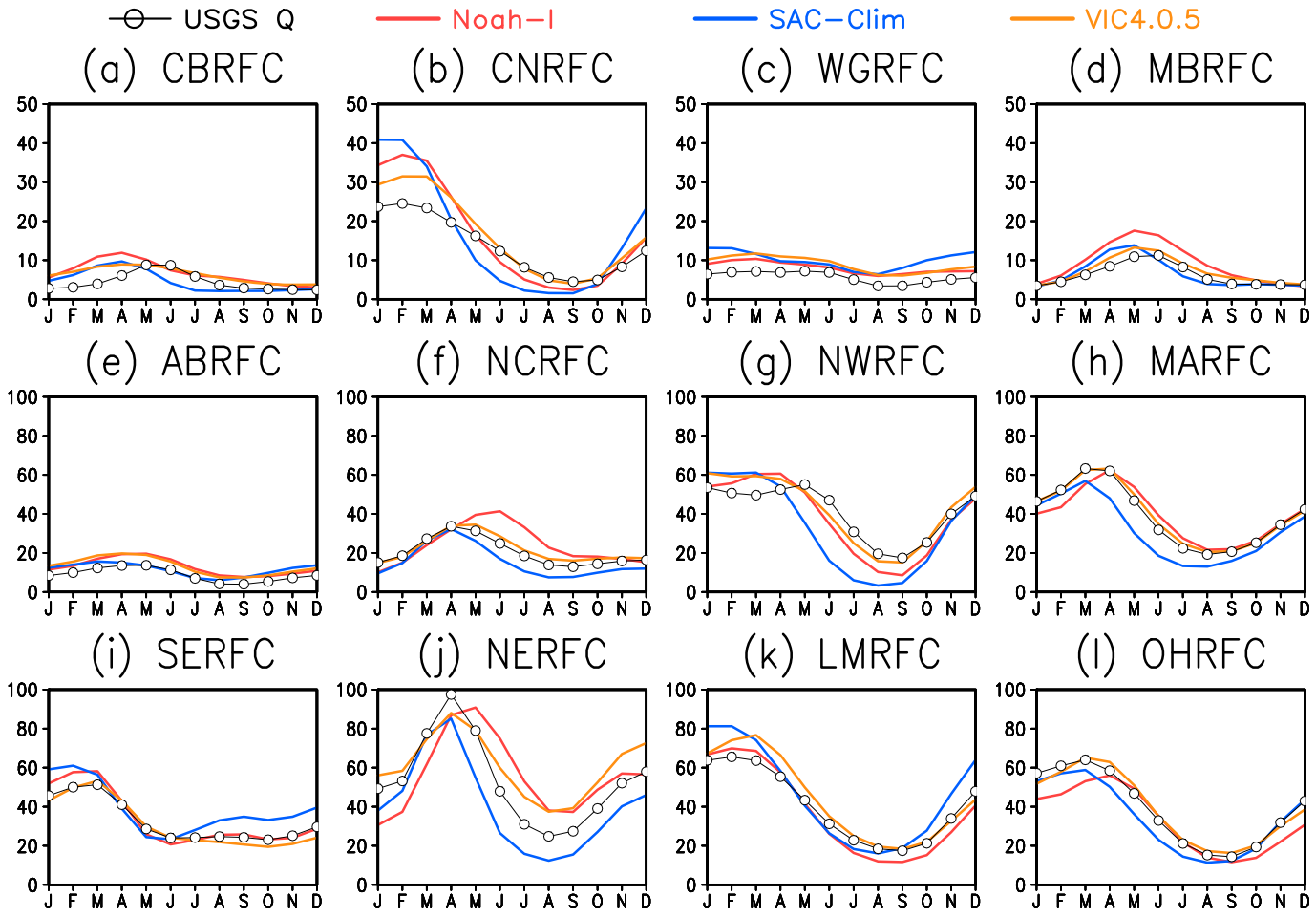

Figure 15. As in Figure 14 except for *Q*. The observation-based reference values (black line with open circles) are from USGS *Q*.

Table 7. Anomaly Correlation (AC) Coefficients Between References and the Models Used in the Research NLDAS-2[a]

	Q			ET			dS/d*t*		
RFC	Noah-I	SAC-Clim	VIC4.0.5	Noah-I	SAC-Clim	VIC4.0.5	Noah-I	SAC-Clim	VIC4.0.5
CBRFC	**0.70**	0.58	0.65	**0.82**	**0.75**	***0.68***	**0.96**	**0.94**	0.97
CNRFC	**0.96**	***0.88***	**0.96**	**0.82**	**0.74**	***0.60***	0.94	**0.96**	***0.95***
WGRFC	***0.85***	***0.87***	***0.87***	***0.85***	0.86	***0.81***	0.97	**0.95**	**0.97**
MBRFC	0.89	***0.84***	**0.91**	**0.83**	***0.69***	0.82	0.97	**0.95**	0.93
ABRFC	***0.90***	***0.91***	**0.95**	0.80	***0.72***	0.75	**0.94**	0.93	***0.93***
NCRFC	**0.88**	0.90	0.93	**0.54**	−0.08	***0.45***	**0.95**	**0.95**	***0.95***
NWRFC	**0.93**	0.89	***0.91***	**0.63**	***0.44***	0.63	**0.95**	**0.94**	***0.95***
MARFC	**0.92**	0.96	**0.96**	0.44	−0.12	0.34	0.90	0.93	**0.95**
SERFC	0.95	**0.93**	**0.96**	***0.26***	−0.04	0.24	**0.97**	0.96	**0.96**
NERFC	**0.86**	**0.90**	***0.93***	**0.26**	−0.20	**0.62**	**0.92**	**0.95**	0.94
LMRFC	***0.96***	0.93	***0.95***	***0.32***	−0.08	**0.28**	0.96	**0.93**	**0.95**
OHRFC	0.87	**0.94**	0.95	0.53	−0.12	**0.49**	**0.92**	**0.95**	**0.94**

[a]The AC value $> |0.12|$ is significant at the 5% significance level for a Student *t* test. The bold values represent improvements, and bold italic values represent deteriorations at the 5% significance level for a two-tailed test when compared with the AC values calculated from the corresponding operational model version listed in Table 4.

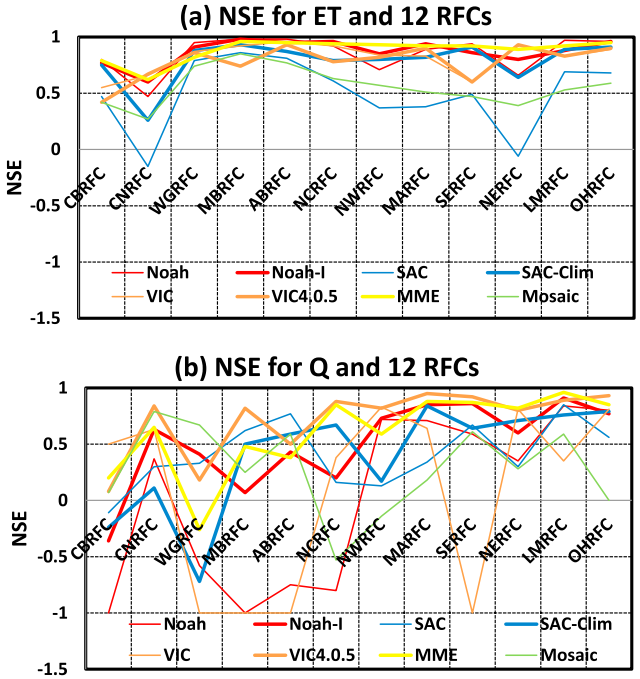

Figure 16. (a) Comparison of Nash-Sutcliffe efficiency (NSE) calculated from the observation-based MTE FLUXNET reference ET and the simulated ET from the operational NLDAS-2 by the LSMs of Noah (thin red line), Mosaic (thin green line), SAC (thin blue line), VIC (thin orange line), and their ensemble mean MME (thick yellow line) and from the research NLDAS-2 by the upgraded LSMs of Noah-I (thick red line), SAC-Clim (thick blue line), and VIC4.0.5 (thick orange line). The RFCs are depicted in order (from left to right) from the driest to the wettest based on the aridity index given in Table 3. (b) Same as Figure 16a except for *Q*, with the observation-based reference *Q* from the USGS.

significant improvements (deteriorations) for VIC4.0.5 when compared to their operational NLDAS-2 counterparts. For the AC of monthly mean *Q* simulated by the three upgraded LSMs, six (three) RFCs show significant improvements (deteriorations) for Noah-I, three (three) RFCs show significant improvements (deteriorations) for SAC-Clim, and five (four) RFCs show significant improvements (deteriorations) for VIC4.0.5. For the AC of monthly mean d*S*/d*t* by the three upgraded LSMs, seven RFCs show significant improvements for the upgraded Noah, nine RFCs show significant improvements for the upgraded SAC-Clim, and none of the three LSMs show obvious deterioration at any RFC. For d*S*/d*t*, the upgraded VIC4.0.5 has five (four) RFCs with significant improvement (deterioration) when compared with the results in the operational NLDAS-2. Overall, Noah-I features significant improvement over more RFCs than do the other two upgraded LSMs. SAC-Clim and VIC4.0.5 have more mixed results as some RFCs have significant improvements, while others show significant deterioration.

As an integrated metric to measure the simulation skill of the LSMs, the Nash-Sutcliffe efficiency (NSE) [*Nash and Sutcliffe*, 1970] is widely used to evaluate simulated streamflow and total runoff [*Lohmann et al*., 2004; *Xia et al*., 2012b] as well as to evaluate simulated ET [*Bhattarai et al*., 2012; *Ershadi et al*., 2014]. The NSE value can vary from $-\infty$ to 1, where 1 corresponds to a perfect match of modeled to observed/reference data. An NSE of 0 indicates that the model simulations are as accurate as the mean of the observed data, whereas an NSE value of less than zero occurs when the observed mean is a better predictor than the model. NSE is calculated as

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{i=N} (S_i - O_i)^2}{\sum_{i=1}^{i=N} (O_i - \bar{O})^2} \tag{3}$$

In equation (3) S_i and O_i are, respectively, simulated and observed/reference variables at the *ith* month, N is the number of total months, and $\bar{S}$ and $\bar{O}$ are their mean values for any given time period.

The NSE values for simulated monthly ET and *Q* are given in Figures 16a and 16b, respectively, for all four LSMs and their ensemble mean (MME) in the operational NLDAS-2 and for all three upgraded LSMs (Noah-I, SAC-Clim, and VIC4.0.5) in the research NLDAS-2. The Mosaic and SAC LSMs in the operational NLDAS-2 have the lowest NSE values of monthly mean ET when compared with the other LSMs in either the operational or research NLDAS-2, suggesting poor ET performance. The other LSMs have comparable performance in NSE values of monthly mean ET, although Noah-I and MME have better performance than Noah, VIC, VIC4.0.5, and SAC-Clim. Overall, the skill of simulated monthly mean ET is good.

For monthly mean *Q*, Noah and VIC in the operational NLDAS-2 have large negative NSE values, suggesting poor performance, in particular for relatively dry RFCs. Generally, the upgraded Noah-I and VIC4.0.5 in the

research NLDAS-2 outperform their operational counterparts for all 12 RFCs. For SAC versus SAC-Clim, the former performs better (worse) than the latter for relatively dry (wet) RFCs. The Mosaic model features better performance over the dry RFCs than over the wet RFCs. Overall for the NSE of *Q*, the performance of VIC4.0.5 is the best, followed by MME, Noah-I, and the others (e.g., Noah, Mosaic, SAC, SAC-Clim, and VIC). Therefore, in general the upgraded LSMs provide better performance than their counterpart versions in the operational NLDAS-2, suggesting that an upgrade of the operational NLDAS-2 is warranted in the near future.

6. Conclusion

The NCEP operational NLDAS-2 system has been developed, implemented, and evaluated against several key reference data sets for the components of the surface water budget at annual and monthly time scales. Additionally, the research NLDAS-2 system is compared with the operational NLDAS-2 system. The two key observation-based gridded reference products used for comparison are the 27 year (1982–2008) (A) monthly FLUXNET analysis fields of evapotranspiration (ET) having 0.5° resolution and (B) monthly USGS analysis fields of runoff (*Q*) having "HUC8" resolution, which is nominally 0.6°. A third set of gridded products used for comparison is the analysis fields of monthly change in total water storage derived from the water budget equation using the observation-based monthly NLDAS-2 precipitation, and the aforementioned FLUXNET ET and USGS *Q*. The fourth set of gridded products used for comparison consists of GRACE-based total water storage anomaly (TWSA) and total water storage change (TWSC). Due to different spatial resolutions among these three reference data sets, a comparison from grid cell to grid cell is very difficult and would suffer from spatial-scale mismatch problems. To reduce the impact of such mismatches, spatial averaging over the area of responsibility of each of the 12 NWS RFCs was used to perform the comparison and evaluation.

This study used the newly released NCDC CONUS monthly precipitation analysis to compare and evaluate NLDAS-2 monthly precipitation. The results show that these two gauge-based precipitation analyses are very similar, with small differences ($<$3%) when averaged over monthly time scales and RFC-basin spatial scales. Furthermore, the mean annual difference between NLDAS-2 precipitation and the sum of FLUXNET ET and USGS *Q* is smaller than 10% for all 12 RFCs, except for ABRFC where there is a 16% difference.

Our evaluation of water budget components focused on: (1) the four land surface models and their multimodel ensemble mean in the NCEP operational NLDAS-2 and (2) the three upgraded LSMs in the research NLDAS-2 system. The comparison was performed for both the mean annual climatology and the annual cycle of monthly mean values over each of the 12 RFCs. The statistical metrics used in this study were anomaly correlation, Nash-Sutcliffe efficiency, bias, and RMSE. For the operational NLDAS-2, Mosaic and SAC overestimated (underestimated) mean annual ET (*Q*), while Noah and VIC underestimated (overestimated) mean annual ET (*Q*) compared to the reference observation-based products. The MME manifested the smallest bias in ET and *Q*.

For all 12 RFCs, all four LSMs broadly capture the annual cycle of the observation-based monthly mean FLUXNET ET and USGS *Q*, as well as the water balance equation-derived change in monthly total surface water storage (dS/dt) for all 12 RFCs, except for (1) ~1 month errors in the month of peak *Q* at some RFCs and (2) the mountainous CNRFC and CBRFC regions, which are characterized by substantial snowpack in the cold season. Analysis of anomaly correlation showed that all four LSMs have quite high AC values for simulated monthly *Q* and dS/dt for all 12 RFCs. Higher (lower) AC values were obtained for the seven relatively dry RFCs (five relatively wet RFCs).

Very importantly, the three upgraded LSMs (i.e., Noah-I, SAC-Clim, and VIC4.0.5) in the research NLDAS-2 system significantly reduce the bias and RMSE of the simulated mean annual ET and *Q* compared to their operational counterparts, with the exception of SAC-Clim *Q* in dry regions. A nearly 50% reduction in bias and RMSE in the research versus operational NLDAS-2 can be found when results are compared over the 12 RFCs. Both the operational and research NLDAS-2 systems use the same forcing fields, so the improvements obtained in the research NLDAS-2 stem solely from the LSM and parameter upgrades.

The comparison of the annual cycle of monthly ET, *Q*, and dS/dt between the research and operational NLDAS-2 shows that the spread and disparity among the three upgraded research LSMs become smaller, and the simulated values are closer to the reference values when compared with that in the operational NLDAS-2. The analysis of statistical performance metrics shows that the overall performance of the annual cycle of monthly mean ET, *Q*, and dS/dt from the three upgraded LSMs is better than their counterparts in

the operational NLDAS-2, although model improvement for the annual cycle of monthly means is not as large as the improvement obtained for the mean annual climatology. Also, in the AC analysis, the upgraded SAC LSM (SAC-Clim) produces a mix of improvement and degradation when compared to the SAC version in the operational NLDAS-2. VIC4.0.5 also shows significant deterioration at some RFCs compared to its operational version. Despite these caveats, the overall impact on NLDAS products of upgrading the three LSMs is a positive one. An upgraded version of the Mosaic LSM was not tested in this study, because the NASA/GSFC Hydrological Sciences Laboratory is focusing on replacing the Mosaic LSM used in NLDAS-2 with a newer NASA/GSFC LSM known as the Catchment Model.

Finally, it is emphasized that although these methodologies have been demonstrated within the NLDAS-2 framework, they are general in form and can be applied to any other land data assimilation systems (LDAS) over the world. The improved scientific understanding of the surface turbulent exchange coefficient calculation under stable boundary conditions (e.g., Noah-I versus Noah) and different exchange coefficient methods (Noah-MP versus Noah, see section 7) are impactful scientific advances for the LSM development community. Additionally, engineering-oriented advances were obtained via reasonable LSM calibration and tuning (VIC4.0.5 versus VIC), and improved surface forcing data (SAC-Clim versus SAC) provide further benefit to the LSM development community. The joint nature of these improvements is critical for enhancing operational/research product quality in various LDAS systems, and these findings shape the future pathway for the next-generation NLDAS system discussed below.

7. Future Pathway

We recognize that various significant weaknesses in simulated products remain for both the operational and research NLDAS-2 systems, although in this study the latter generally performed better than the former. We found that two challenging RFCs are the CNRFC and CBRFC where all four LSMs fail to acceptably simulate the annual cycle of the monthly mean FLUXNET-observed ET and USGS-observed *Q*. These weaknesses will be scrutinized and addressed by a collaborative effort between the EMC NLDAS team, its external partners, and the LSM development community for the Noah, SAC, VIC, and Catchment LSMs. These efforts will include reexamination of (1) LSM physical processes (e.g., snowpack treatment, PET estimation, surface drag coefficient, topographic effects, and simulation of ET and *Q* during summer dry periods) and (2) the algorithms and data sources for NLDAS-2 forcing fields (e.g., snowfall and rainfall partitioning, downward radiation, and surface air temperature).

Recently, the NWS/NWC upgraded the SAC model to a new SAC-HT-ET version by (1) adding a soil heat transfer capacity across different soil layers (e.g., soil temperature calculation) [*Koren et al*., 2014], (2) adding surface energy balance processes including the calculation of ET [*Koren et al*., 2007, 2010], and (3) conducting CONUS-wide testing of the model. This SVAT-like model is being tested using NLDAS-2 forcing in the Land Information System (LIS) framework at the NASA/GSFC Hydrological Sciences Lab. In addition, the EMC land team is collaborating with our NLDAS partners (e.g., NCAR, NASA, and NWC) to test Noah-MP, CLM4, and CLSM-F2.5 and is testing an irrigation module in the NLDAS-2 LSMs. Noah-MP, CLM4, and CLSM-F2.5 contain groundwater modules and so are able to simulate variation in the groundwater table and the exchange of water between groundwater and deep soil layers. After these upgrades are completed, the performance of Noah/Noah-MP, CLM4, and CLSM-F2.5 will be comprehensively reevaluated to study the impact of irrigation and groundwater on ET using the same data sets and framework as used in this study.

A preliminary result here is that we compared seasonal variations of the total runoff simulated by Noah-MP and Noah-I with USGS-observed values at 12 RFCs (Figure 17). In this study, we use the same Noah-MP version (EXP6) as used in *Niu et al*. [2011]. It includes dynamic vegetation, groundwater table, multilayer snow model, and other physical processes' updates. The results show that Noah-MP improves total runoff simulation in terms of timing and/or amplitude for CBRFC, CNRFC, MBRFC, NCRFC, and NERFC, although it overestimates (underestimates) the runoff simulation in WGRFC and SERFC (OHRFC). For the noted runoff improvement over the cold RFCs, the improvement is mainly due to the fact that Noah-MP more reasonably simulates the seasonal cycle of snowmelt by replacing the operational surface exchange coefficient scheme [*Chen et al*., 1997] with a scheme based on Monin-Obukhov (M-O) similarity theory [*Monin and Obukhov*, 1954], although introduction of a multilayer structure in the snowpack model also plays a moderate role [*Niu et al*., 2011].

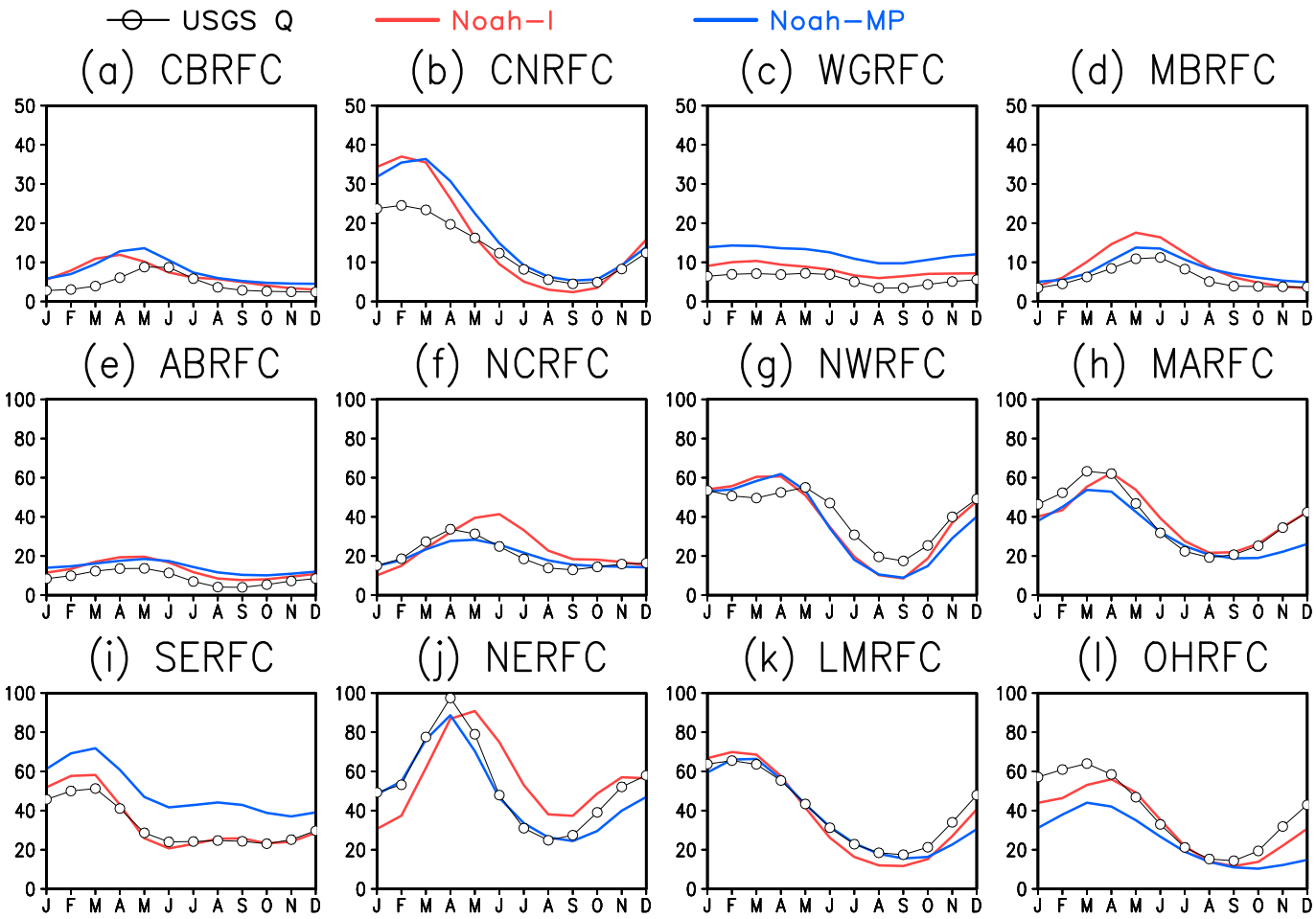

Figure 17. For each of the 12 RFCs, comparison of the 27 year (1982–2008) mean annual cycle of monthly mean *Q* (unit: mm/month) of the observation-based USGS reference (black line with open circles) with that simulated in the research NLDAS-2 by Noah-I (red) and Noah-MP (blue).

As indicated by *Yang et al*. [2011], the M-O scheme produces smaller surface turbulent exchange coefficient (CH) values in the cold season, which reduces large sublimation values found in the Noah LSM [*Slater et al*., 2007]. We now recognize, however, that the stable condition CH constraints applied in both Noah 2.8 [*Livneh et al*., 2010] and Noah-I [*Xia et al*., 2014c], which both manually constrain the CH value by imposing a lower bound for either all stable (Noah 2.8) or all snow-covered stable (Noah-I) surface layer conditions is only an intermediate solution. Although these stable CH constraints partially solve the large sublimation and early snowmelt issues found in *Slater et al*. [2007], these same constraints degrade the overall total runoff and evapotranspiration simulations, as described in this study, as well as degrade the simulations of other variables, as shown in the previous study of *Xia et al*. [2014c].

For the relatively warm climatic conditions of the WGRFC and SERFC, Noah-MP overestimates total runoff when compared with USGS-observed values. This overestimation also occurs in the previous study of *Yang et al*. [2011]. For example, Noah-MP also overestimates total runoff for the tropical river basin in the Congo and Tocantins river basin in Brazil. A potential reason is that the smaller CH values generated by the M-O scheme leads to smaller ET values, which in turn yields larger runoff. This hypothesis requires more investigation in the future. Although vegetation and groundwater dynamics are not major factors in this study when total runoff simulations are investigated at basin scale, their impact on water cycles still needs to be investigated in the future. In line with this direction, many sensitivity tests and comparative analyses for Noah-MP, as well as the other upgraded land surface models discussed below, are underway in the NASA NLDAS Science Testbed.

We also recognize that both the NCEP operational NLDAS-2 and the EMC research NLDAS-2 are not actual land data assimilation systems per se, because there is no assimilation of remotely sensed land surface states

such as soil moisture and snowpack. To date, the data assimilation character of NLDAS-1 and NLDAS-2 has been based on the various EMC data assimilation systems that produce the NLDAS forcing fields (e.g.; that of the NCEP Regional Reanalysis [*Mesinger et al*., 2006]) and the bias corrections applied to the latter [*Cosgrove et al*., 2003]. Striving to add land data assimilation, the NASA GSFC Hydrological Sciences Lab is currently collaborating with NCEP EMC to make NASA's LIS assimilation suite the overarching software suite used for the research (and eventually into operations) NLDAS system. LIS contains multiple data assimilation modules. The switch from individual non-data-assimilation model drivers to a unified LIS software framework in the research NLDAS-2 will greatly facilitate the eventual assimilation of soil moisture [*Kumar et al*., 2014], snowpack such as snow depth and snow cover [*Kumar et al*., 2015; *Liu et al*., 2015], total water storage change (retrieved from GRACE [*Swenson and Wahr*, 2002; *Wahr et al*., 2004]), and other variables, such as satellite-retrieved brightness temperatures and streamflow observed at basin outlets from the USGS.

Specifically, at NASA/GSFC [*Kumar et al*., 2014] a preliminary test in LIS has demonstrated actual land data assimilation to improve soil moisture, snowpack, and total runoff simulations in Noah Version 3.3. This collaboration is an ongoing project supported by the NOAA Climate Program Office to develop the next-generation NLDAS-3 system not only by applying the LIS assimilation infrastructure but also by upgrading all NLDAS LSMs to their latest versions and adding ground water and irrigation modules. Test runs of Noah LSM Version 3.3, Noah-3.6, and the CLSM-F2.5 have been completed at NASA/GSFC. These test runs are being evaluated against in situ observations and satellite retrievals and will soon be transitioned to NCEP/EMC to perform preoperational test runs in the research NLDAS-2. In addition, SAC-HT-ET (SAC Heat Transfer and Evapotranspiration model, the latest version of SAC [*Koren et al*., 2007, 2010]) and VIC4.1.2 are also running at NASA/GSFC and will be evaluated and transitioned to NCEP/EMC after sufficient assessments have been completed.

Acknowledgments

NLDAS-1 research activities have been supported by the NOAA Office of Global Programs Global Energy and Water Cycle Experiment (GEWEX) Americas Prediction Project (GAPP) and the NASA Terrestrial Hydrology Program. NLDAS-2 research and operational transition activities have been supported by the NOAA Climate Program Office (CPO) Climate Prediction Program of the Americas (CPPA) and Modeling Analysis, Predictions, and Projections (MAPP). We acknowledge Dr. David Wolock, who helped us create the HUC8 index mask file for the NLDAS-2 grid. We also acknowledge Fanglin Yang and Hong Guan from EMC and three anonymous reviewers whose review and comments greatly improved the quality of the manuscript. All data including NLDAS-2 products, USGS runoff, GRACE-observed TWSA and TWSC, and gridded FLUXNET data can be freely accessed via public websites as described in the text.

References

Ashfaq, M., S. Ghosh, S.-C. Kao, L. C. Bowling, P. Mote, D. Touma, S. A. Rauscher, and N. S. Diffenbaugh (2013), Near-term acceleration of hydroclimatic change in the western U.S, *J. Geophys. Res. Atmos.*, *118*, 10,676–10,693, doi:10.1002/jgrd.50816.

Baldocchi, D. D., et al. (2001), FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, *Bull. Am. Meteorol. Soc.*, *82*(11), 2415–2434.

Bhattarai, N., M. Dougherty, L. J. Marzen, and L. Kalin (2012), Validation of evaporation estimates from a modified surface energy balance algorithm for land (SEBAL) model in the south-eastern United States, *Remote Sens. Lett.*, *3*, 511–519.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A generalized streamflow simulation system: Conceptual models for digital computer, technical report, Joint Fed.-State River Forecast Cent., U.S. Natl. Weather Serv. And Calif. Dep. of Water Resour., Sacramento, Calif.

Cai, X. T., Z.-L. Yang, Y. Xia, M. Huang, H. Wei, R. Leung, and M. B. Ek (2014), Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS Testbed, *J. Geophys. Res. Atmos.*, *119*, 13,751–13,770, doi:10.1002/2014JD022113.

Chen, F., Z. Janjic, and K. E. Mitchell (1997), Impact of atmospheric surface layer parameterizations in the new land surface scheme of the NCEP mesoscale Eta model, *Boundary Layer Meteorol.*, *85*, 391–421, doi:10.1023/A:1000531001463.

Cosgrove, B. A., et al. (2003), Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, *J. Geophys. Res.*, *108*(D22), 8842, doi:10.1029/2002JD003118.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical-topographic model for mapping climatological precipitation over mountainous terrain, *J. Appl. Meteorol.*, *33*, 140–158.

Dooge, J. C. I. (1997), Scale problems in hydrology, in *Reflections in Hydrology*, edited by N. Buras, pp. 85–143, AGU, Washington, D. C.

Ducharne, A., R. D. Koster, M. J. Suarez, M. Stieglitz, and P. Kumar (2000), A catchment-based approach to modeling land surface processes in a GCM, Part 2, Parameter estimation and model demonstration, *J. Geophys. Res.*, *105*, 24,823–24,838, doi:10.1029/2000JD900328.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rodgers, P. Grunman, V. Koren, G. Gayno, and J. D. Tarpley (2003), Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, *108*(D22), 8851, doi:10.1029/2002JD003296.

Ek, M. B., et al. (2011), North American Land Data Assimilation System Phase 2 (NLDAS-2): Development and applications, *GEWEX News*, *21*(2), 6–7.

Ershadi, A., M. F. McCabe, J. P. Evans, N. W. Chaney, and E. F. Wood (2014), Multi-site evaluation of terrestrial evaporation models using FLUXNET data, *Agric. For. Meteorol.*, *187*, 46–61.

Gandin, L. (1963), Objective analysis of meteorological fields, Gidro-meteorologichoskoe Isdatel'stvo, Leningrad, translated from Russian, Israel Program for Scientific Translation, Jerusalem, *Q. J. R. Meteorol. Soc.*, *92*, 447, doi:10.1002/qj.49709239320.

Gent, P. R., S. G. Yeager, R. B. Neale, S. Levis, and D. A. Bailey (2010), Improvements in a half degree atmosphere/land version of the CCSM, *Clim. Dyn.*, *34*(6), 819–833, doi:10.1007/s00382-009-0614-8.

Getirana, A. C. V., et al. (2014), Water balance in Amazon basin from a land surface model ensemble, *J. Hydrometeorol.*, *15*, 2586–2614, doi:10.1175/JHM-D-14-0068.1.

Hobbins, M. T., J. A. Ramı´rez, and T. C. Brown (2001), The complementary relationship in estimation of regional evapotranspiration: An enhanced advection-aridity model, *Water Resour. Res.*, *37*(5), 1389–1403, doi:10.1029/2000WR900359.

Hutchinson, M. F. (1995), Interpolating mean rainfall using thin plate smoothing splines, *Int. J. GIS*, *9*, 305–403.

Jarvis, P. G. (1976), The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field, *Philos. Trans. R. Soc. London Ser. B.*, *273*, 593–610.

Jiménez, C., et al. (2011), Global intercomparison of 12 land surface heat flux estimates, *J. Geophys. Res.*, *116*, D02102, doi:10.1029/2010JD014545.

Jung, M., M. Reichstein, and A. Bondeau (2009), Towards global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model, *Biogeosciences*, *6*, 2001–2013.

Jung, M., et al. (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res.*, *116*, G00J07, doi:10.1029/2010JG001566.
Koren, V. I., M. Smith, and Z. Cui (2014), Physically-based modifications to the Sacramento Soil Moisture Accounting model. Part A: Modeling the effects of frozen ground on the runoff generation process, *J. Hydrol.*, *519*, 3475–3491, doi:10.1016/j.jhydrol.2014.03.004.
Koren, V., M. Smith, Z. Cui, and B. Cosgrove (2007), Physically-based modifications to the Sacramento Soil Moisture Accounting Model: Modeling the effects of frozen ground on the rainfall-runoff process, NOAA Tech. Rep., *52*, 47 pp., National Weather Service, Silver Spring, Md. [Available at http://www.nws.noaa.gov/oh/hrl/hsmb/hydrology/PBE_SAC-SMA/index.html.]
Koren, V., M. Smith, Z. Cui, B. Cosgrove, K. Werner, and R. Zamora (2010), Modification of Sacramento Soil Moisture Accounting Heat Transfer Component (SAC-HT) for enhanced evapotranspiration, *NOAA Tech. Rep.*, *53*, 72 pp., National Weather Service, Silver Spring, Md. [Available at http://www.nws.noaa.gov/oh/hrl/hsmb/hydrology/PBE_SAC-SMA/index.html.]
Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar (2000), A catchment-based approach to modeling land surface processes in a GCM, Part 1, Model structure, *J. Geophys. Res.*, *105*, 24,809–24,822, doi:10.1029/2000JD900327.
Koster, R., and M. Suarez (1994), The components of the SVAT scheme and their effects on a GCM's hydrological cycle, *Adv. Water Resour.*, *17*, 61–78.
Kumar, S. V., et al. (2006), Land Information System—An interoperable framework for high resolution land surface modeling, *Environ. Modell. Software*, *21*, 1402–1415.
Kumar, S. V., et al. (2014), Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation, *J. Hydrometeorol.*, *15*, 2446–2469.
Kumar, S. V., C. D. Peters-Lidard, K. R. Arsenault, A. Getirana, D. Mocko, and Y. Liu (2015), Quantifying the added value of snow cover area observations in passive microwave snow depth data assimilation, *J. Hydrometeorol.*, *16*, 1736–1741, doi:10.1175/JHM-D-15-0021.1.
Landerer, F. W., and S. C. Swenson (2012), Accuracy of scaled GRACE terrestrial water storage estimates, *Water Resour. Res.*, *48*, W045531, doi:10.1029/2011WR011453.
Lawrence, D. M., K. W. Oleson, M. G. Flanner, C. G. Fletcher, P. J. Lawrence, S. Levis, S. C. Swenson, and G. B. Bonan (2012), The CCSM4 land simulations, 1850–2005: Assessment of surface climate and new capabilities, *J. Clim.*, *25*(7), 2240–2260, doi:10.1175/jcli-d-11-00103.1.
Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A simple hydrologically based model of land surface water and energy fluxes for GCMs, *J. Geophys. Res.*, *99*, 14,415–14,428, doi:10.1029/94JD00483.
Liu, Y., C. Peters-Lidard, S. Kumar, K. Arsenault, and D. Mocko (2015), Blending satellite-based snow depth products with in situ observations for streamflow predictions in the Upper Colorado River Basin, *Water Resour. Res.*, *51*, 1182–1202, doi:10.1002/2014WR016606.
Livneh, B., Y. Xia, K. E. Mitchell, M. B. Ek, and D. P. Lettenmaier (2010), Noah LSM Snow Model diagnostics and enhancements, *J. Hydrometeorol.*, *11*, 721–738.
Lohmann, D., et al. (2004), Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project, *J. Geophys. Res.*, *109*, D07S91, doi:10.1029/2003JD003517.
Long, D., B. R. Scanlon, L. Longuevergne, A. Y. Sun, D. N. Fernando, and H. Save (2013), GRACE satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas, *Geophys. Res. Lett.*, *40*, 3395–3401, doi:10.1002/grl.50655.
Long, D., L. Longuevergne, and B. R. Scanlon (2014), Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites, *Water Resour. Res.*, *50*, 1131–1151, doi:10.1002/2013WR014581.
Mesinger, F., et al. (2006), North American regional reanalysis, *Bull. Am. Meteorol. Soc.*, *87*, 343–360.
Mitchell, K. E., et al. (2004), The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, *J. Geophys. Res.*, *109*, D07S90, doi:10.1029/2003JD003823.
Mo, K. C., L. N. Long, Y. Xia, S. K. Yang, J. E. Schemm, and M. Ek (2011), Drought indices based on the climate forecast system reanalysis and ensemble NLDAS, *J. Hydrometeorol.*, *12*, 181–205.
Monin, A. S., and A. M. Obukhov (1954), Basic laws of turbulent mixing in the surface layer of the atmosphere, *Tr. Akad. Nauk SSSR Geofiz. Inst.*, *24*, 163–187.
Mueller, B., et al. (2011), Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations, *Geophys. Res. Lett.*, *38*, L06402, doi:10.1029/2010GL046230.
Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models: Part A—A discussion of principles, *J. Hydrol.*, *10*, 282–290.
Niu, G.-Y., et al. (2011), The community Noah land surface model with multi-physics options, part 1: Model descriptions and evaluation with local-scale measurements, *J. Geophys. Res.*, *116*, D12109, doi:10.1029/2010JD015139.
Oubeidillah, A. A., S.-C. Kao, M. Ashfaq, B. S. Naz, and G. Tootle (2014), A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US, *Hydrol. Earth Syst. Sci.*, *18*, 67–84.
Peters-Lidard, C. D., et al. (2007), High-performance Earth system modeling with NASA/GSFC's Land Information System, *Innovations Syst. Software Eng.*, *3*, 157–165.
Peters-Lidard, C. D., S. V. Kumar, D. M. Mocko, and Y. Tian (2011), Estimating evapotranspiration with land data assimilation systems, *Hydrol. Processes*, *25*, 3979–3992.
Robock, A., et al. (2003), Evaluation of the North American Land Data Assimilation System over the southern Great Plains during the warm season, *J. Geophys. Res.*, *108*(D22), 8846, doi:10.1029/2002JD003245.
Royer, A., and S. Poirier (2010), Surface temperature spatial and temporal variations in North America from homogenized satellite SMMR-SSM/I microwave measurements and reanalysis for 1979–2008, *J. Geophys. Res.*, *115*, D08110, doi:10.1029/2009JD0127.
Sakumura, C., S. Bettadpur, and S. Bruinsma (2014), Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models, *Geophys. Res. Lett.*, *41*, 1389–1397, doi:10.1002/2013GL058632.
Sankarasubramanian, A., and R. M. Vogel (2002), Annual hydroclimatology of the United States, *Water Resour. Res.*, *38*(6), 1083, doi:10.1029/2001WR000619.
Saville, D. J. (1990), Multiple comparison procedures: The practical solution, *Am. Stat.*, *44*, 174–180.
Schaake, J. C., et al. (2004), An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS), *J. Geophys. Res.*, *109*, D01S90, doi:10.1029/2002JD003309.
Sellers, P., Y. Mintz, Y. Sud, and A. Dalcher (1986), A simple Biosphere model (SiB) for use within general circulation models, *J. Atmos. Sci.*, *43*, 505–531.
Sharif, H. O., and F. L. Ogden (2014), Mass-conserving remapping of radar data onto two-dimensional cartesian coordinates for hydrologic applications, *J. Hydrometeorol.*, *15*, 2190–2202, doi:10.1175/JHM-D-14-0058.1.
Slater, A. G., T. J. Bohn, J. L. McCreight, M. C. Serreze, and D. P. Lettenmaier (2007), A multimodel simulation of pan-Arctic hydrology, *J. Geophys. Res.*, *112*, G04S45, doi:10.1029/2006JG000303.
Steiger, J. H. (1980), Tests for comparing elements of a correlation matrix, *Psychol. Bull.*, *87*, 245–251.

Strassberg, G., B. R. Scanlon, and D. Chambers (2009), Evaluation of ground water storage monitoring with GRACE satellite: Case study of the High Plains aquifer, central United States, *Water Resour. Res.*, *45*, W05410, doi:10.1029/2008WR006892.
Swenson, S., and J. Wahr (2002), Methods for inferring regional surface-mass anomalies from Gravity Recovery and Climate Experiment (GRACE) measurements of time-variable gravity, *J. Geophys. Res.*, *107*(B9), 2193, doi:10.1029/2001JB000576.
Swenson, S., P. J.-F. Yeh, J. Wahr, and J. Famiglietti (2006), A comparison of terrestrial water storage variations from GRACE with in situ measurements from Illinois, *Geophys. Res. Lett.*, *33*, L16401, doi:10.1029/2006GL026962.
Troy, T. J., E. F. Wood, and J. Sheffield (2008), An efficient calibration method for continental-scale land surface modeling, *Water Resour. Res.*, *44*, W09411, doi:10.1029/2007WR006513.
Velpuri, N. M., and G. B. Senay (2013), Analysis of long-term trends (1950–2009) in precipitation, runoff and runoff coefficient in major urban watersheds in the United States, *Environ. Res. Lett.*, *8*, doi:10.1088/1748-9236/8/024020.
Velpuri, N. M., G. B. Senay, R. K. Singh, S. Bohms, and J. P. Verdin (2013), A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET, *Remote Sens. Environ.*, *139*, 35–49.
Vose, R. S., et al. (2012), NOAA's merged land–ocean surface temperature analysis, *Bull. Am. Meteorol. Soc.*, *93*, 1677–1685.
Vose, R. S., S. Applequist, M. Squires, I. Durre, M. J. Menne, C. N. Williams, C. Fenimore, K. Gleason, and D. Arndt (2014), Improved historical temperature and precipitation time series for U.S. climate divisions, *J. Appl. Meteorol. Climatol.*, *53*, 1232–1251.
Wahr, J., S. Swenson, V. Zlotnicki, and I. Velicogna (2004), Time-variable gravity from GRACE: First results, *Geophys. Res. Lett.*, *31*, L11501, doi:10.1029/2004GL019779.
Wahr, J., S. Swenson, and I. Velicogna (2006), Accuracy of GRACE mass estimates, *Geophys. Res. Lett.*, *33*, L06401, doi:10.1029/2005GL025305.
Wei, H., Y. Xia, K. E. Mitchell, and M. B. Ek (2013), Improvement of the Noah land surface model for warm season processes: Evaluation of water and energy flux simulation, *Hydrol. Processes*, *27*, 297–303.
Wood, E. F., D. P. Lettenmaier, X. Liang, B. Nijssen, and S. W. Wetzel (1997), Hydrological modeling of continental-scale basins, *Annu. Rev. Erath Planet Sci.*, *25*, 279–300.
Xia, Y., et al. (2012a), Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, *J. Geophys. Res.*, *117*, D03109, doi:10.1029/2011JD016048.
Xia, Y., et al. (2012b), Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow, *J. Geophys. Res.*, *117*, D03110, doi:10.1029/2011JD016051.
Xia, Y., M. Ek, H. Wei, and J. Meng (2012c), Comparative analysis of relationships between NLDAS-2 forcings and model outputs, *Hydrol. Processes*, *26*, 467–474.
Xia, Y., M. Ek, J. Sheffield, B. Livneh, M. Huang, H. Wei, S. Feng, L. Luo, J. Meng, and E. Wood (2013), Validation of Noah-simulated soil temperature in the North American Land Data Assimilation System Phase 2, *J. Appl. Meteorol. Climatol.*, *52*, 455–471.
Xia, Y., J. Sheffield, M. B. Ek, J. Dong, N. Chaney, H. Wei, J. Meng, and E. F. Wood (2014a), Evaluation of multi-model simulated soil moisture in NLDAS-2, *J. Hydrol.*, *512*, 107–125.
Xia, Y., M. T. Hobbins, Q. Mu, and M. B. Ek (2014b), Evaluation of NLDAS-2 evapotranspiration against tower flux site observations, *Hydrol. Processes*, doi:10.1002/hyp.10299.
Xia, Y., C. D. Peter-Lidard, M. Huang, H. Wei, and M. Ek (2014c), Improved NLDAS-2 Noah-simulated hydrometeorological products with an interim run, *Hydrol. Processes*, doi:10.1002/hyp.10190.
Xia, Y., M. B. Ek, Y. Wu, T. W. Ford, and S. M. Quiring (2015a), Comparison of NLDAS-2 Simulated and NASMD observed daily soil moisture. Part I: Comparison and analysis, *J. Hydrometeorol.*, *16*, 1981–2000, doi:10.1175/JHM-D-14-0096.1.
Xia, Y., B. A. Cosgrove, K. E. Mitchell, C. D. Peters-Lidard, M. B. Ek, S. Kumar, D. Mocko, and H. Wei (2015b), Basin-scale assessment of the land surface energy budget in the National Centers for Environmental Prediction Operational and Research NLDAS-2 systems, *J. Geophys. Res. Atmos.*, doi:10.1029/2015JD023889.
Yang, Z.-L., et al. (2011), The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins, *J. Geophys. Res.*, *116*, D12110, doi:10.1029/2010JD015140.